

The American Economic Review

COK-401166-101-P007831
ARTICLES

Review Studies

- GERARD DEBREU The Mathematization of Economic Theory
- DAVID F. HENDRY AND NEIL R. ERICSSON
An Econometric Analysis of U.K. Money Demand in
Monetary Trends in the United States and the United Kingdom
by Milton Friedman and Anna J. Schwartz
- MILTON FRIEDMAN AND ANNA J. SCHWARTZ
Alternative Approaches to Analyzing Economic Data
- LAWRENCE M. AUSUBEL
The Failure of Competition in the Credit Card Market
- JESS BENHABIB AND BOYAN JOVANOVIĆ
Externalities and Growth Accounting
- SHARON G. LEVIN AND PAULA E. STEPHAN
Research Productivity Over the Life Cycle: Evidence for
Academic Scientists
- WILLIAM G. GALE Economic Effects of Federal Credit Programs
- JOHN DOUGLAS WILSON
Optimal Public Good Provision with Limited
Lump-Sum Taxation
- ALAN J. AUERBACH
Retrospective Capital Gains Taxation
- RICHARD ARNOTT AND JOSEPH E. STIGLITZ
Moral Hazard and Nonmarket Institutions: Dysfunctional
Crowding Out or Peer Monitoring?
- JULIO J. ROTEMBERG
A Theory of Inefficient Intrafirm Transactions
- MARK STEGEMAN
Advertising in Competitive Markets
- KYLE BAGWELL AND MICHAEL H. RIORDAN
High and Declining Prices Signal Product Quality
- RAQUEL FERNANDEZ AND JACOB GLÄZER
Striking for a Bargain Between Two Completely
Informed Agents
- ROBERT FORSYTHE, JOHN KENNAN, AND BARRY SOPHER
An Experimental Analysis of Strikes in Bargaining Games
with One-Sided Private Information

101

SHORTER PAPERS: R. J. Caballero; W. Lord and P. Rangazas; K. reidman and B. Dowd; D. Fullerton;
M. S. McPherson and M. O. Schapiro; C. J. Ruhm; J. H. Bergstrand; B. Lind and C. R. Plott; R. G. Hansen and
J. R. Lott, Jr.; J. H. Kagel and D. Levin; D. Levin and J. L. Smith; T. Cowell and A. Glazer; R. J. Barro and
P. M. Romer; Y. Fu

MARCH 1991

THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

Officers

President

THOMAS C. SCHELLING
University of Maryland

President-elect

WILLIAM VICKREY
Columbia University

Vice-Presidents

HENRY J. AARON
The Brookings Institution
CLAUDIA D. GOLDIN
Harvard University

Secretary-Treasurer

C. ELTON HINSHAW
Vanderbilt University

Editor of The American Economic Review

ORLEY C. ASHENFELTER
Princeton University

Editor of The Journal of Economic Literature

JOHN PENCEVEL
Stanford University

Editor of The Journal of Economic Perspectives

JOSEPH E. STIGLITZ
Stanford University

Executive Committee

Elected Members of the Executive Committee

STANLEY FISCHER
Massachusetts Institute of Technology
LAWRENCE H. SUMMERS
Harvard University
GREGORY C. CHOW
Princeton University
SUSAN ROSE-ACKERMAN
Yale University
MICHAEL J. PIORE
Massachusetts Institute of Technology
GAVIN WRIGHT
Stanford University

EX OFFICIO Member

GERARD DEBREU
University of California-Berkeley

•Printed at Banta Company, Menasha, Wisconsin.

•Copyright © American Economic Association 1991. All rights reserved.

•No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

Correspondence relating to advertising, business matters, permissions to quote, subscriptions, and changes of address, should be sent to the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Change of address notice must be received at least six (6) weeks prior to the publication month. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

THE AMERICAN ECONOMIC REVIEW (ISSN 0002-8282), March 1991, Vol. 81, No. 1, is published five times a year (March, May, June, September, December) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Annual fees for regular membership, of which 30 percent is for a year's subscription to this journal, are: \$44.00, \$52.80, or \$61.60 depending on income. A membership also includes the *Journal of Economic Literature* and the *Journal of Economic Perspectives*. In countries other than the U.S.A., add \$16.00 for extra postage. Second-class postage paid at Nashville, TN and at additional mailing offices. POSTMASTER: Send address changes to the *American Economic Review*, 2014 Broadway, Suite 305, Nashville, TN 37203.

THE AMERICAN ECONOMIC REVIEW

Editor

ORLEY ASHENFELTER

Co-Editors

ROBERT H. HAVEMAN

BENNETT T. McCALLUM

PAUL R. MILGROM

Acting Co-Editor

JOHN Y. CAMPBELL

Production Editor

J. DAVID BALDWIN

Board of Editors

GEORGE A. AKERLOF

JAMES E. ANDERSON

TIMOTHY F. BRESNAHAN

HENRY S. FARBER

MARJORIE A. FLAVIN

ROBERT P. FLOOD

CLAUDIA D. GOLDIN

JO ANNA GRAY

REUBEN GRONAU

DANIEL S. HAMERMESH

ROBERT J. HODRICK

KEVIN D. HOOVER

KENNETH L. JUDD

JOHN H. KAGEL

JOHN F. KENNAN

EDGAR O. OLSEN

JOHN G. RILEY

RICHARD ROLL

THOMAS ROMER

DAVID E. M. SAPPINGTON

ROBERT S. SMITH

BARBARA J. SPENCER

RICHARD TRESCH

HAL R. VARIAN

KENNETH WEST

JOHN D. WILSON

LESLIE YOUNG

March 1991

VOLUME 81, NUMBER 1

391
001

Articles

- The Mathematization of Economic Theory
Gerard Debreu 1
- An Econometric Analysis of U.K. Money Demand in *Monetary Trends in the United States and the United Kingdom* by Milton Friedman and Anna J. Schwartz
David F. Hendry and Neil R. Ericsson 8
- Alternative Approaches to Analyzing Economic Data
Milton Friedman and Anna J. Schwartz 39
- The Failure of Competition in the Credit Card Market
Lawrence M. Ausubel 50
- Externalities and Growth Accounting
Jess Benhabib and Boyan Jovanovic 82
- Research Productivity Over the Life Cycle: Evidence for Academic Scientists
Sharon G. Levin and Paula E. Stephan 114
- Economic Effects of Federal Credit Programs
William G. Gale 133
- Optimal Public Good Provision with Limited Lump-Sum Taxation
John Douglas Wilson 153
- Retrospective Capital Gains Taxation
Alan J. Auerbach 167
- Moral Hazard and Nonmarket Institutions: Dysfunctional Crowding Out or Peer Monitoring?
Richard Arnott and Joseph E. Stiglitz 179
- A Theory of Inefficient Intrafirm Transactions
Julio J. Rotemberg 191
- Advertising in Competitive Markets
Mark Stegeman 210
- High and Declining Prices Signal Product Quality
Kyle Bagwell and Michael H. Riordan 224
- Striking for a Bargain Between Two Completely Informed Agents
Raquel Fernandez and Jacob Glazer 240
- An Experimental Analysis of Strikes in Bargaining Games with One-Sided Private Information
Robert Forsythe, John Kennan, and Barry Sopher 253

Shorter Papers

On The Sign of the Investment–Uncertainty Relationship Savings and Wealth in Models with Altruistic Bequests	<i>Ricardo J. Caballero</i>	279
	<i>William Lord and Peter Rangazas</i>	289
A New Estimate of the Welfare Loss of Excess Health Insurance	<i>Roger Feldman and Bryan Dowd</i>	297
Reconciling Recent Estimates of the Marginal Welfare Cost of Taxation	<i>Don Fullerton</i>	302
Does Student Aid Affect College Enrollment? New Evidence on a Persistent Controversy	<i>Michael S. McPherson and Morton Owen Schapiro</i>	309
Are Workers Permanently Scarred by Job Displacements?	<i>Christopher J. Ruhm</i>	319
Structural Determinants of Real Exchange Rates and National Price Levels: Some Empirical Evidence	<i>Jeffrey H. Bergstrand</i>	325
The Winner's Curse: Experiments with Buyers and with Sellers	<i>Barry Lird and Charles R. Plott</i>	335
The Winner's Curse and Public Information in Common Value Auctions:		
Comment	<i>Robert G. Hansen and John R. Lott, Jr.</i>	347
Reply	<i>John H. Kagel and Dan Levin</i>	362
Some Evidence on the Winner's Curse:		
Comment	<i>Dan Levin and James L. Smith</i>	370
Ski-Lift Pricing, with Applications to Labor and Other Markets:		
Comment	<i>Tyler Cowen and Amihai Glazer</i>	376
Reply	<i>Robert J. Barro and Paul M. Romer</i>	378
A Model of Housing Tenure Choice:		
Comment	<i>Yuming Fu</i>	381

P 7831

Erratum

Cooperation, Harassment, and Involuntary Unemployment	<i>Ernst Fehr</i>	384
---	-------------------	-----

The following Statement of Ownership, Management, and Circulation is provided in accordance with the requirements, as contained in 39 U.S. Code 3685. The *American Economic Review* is owned, managed, and published by the American Economic Association, a nonprofit scientific organization, located at 2014 Broadway, Suite 305, Nashville, Davidson County, Tennessee 37203-2418. The Editor is Professor Orley Ashenfelter, *American Economic Review*, 209 Nassau Street, Princeton, NJ 08542-4607. During the preceding 12 months the average number of copies printed for each issue was 27,949; the average paid circulation, 25,960; the average free distribution, 90; the average number of copies distributed, 26,050. Corresponding figures for the issue published nearest to the filing date: total number of copies printed, 28,000; total paid circulation, 26,129; total free distribution, 90; total distribution, 26,219.

391
001

•Submit manuscripts (4 copies), 50 pages maximum, single-sided, double-spaced, to:

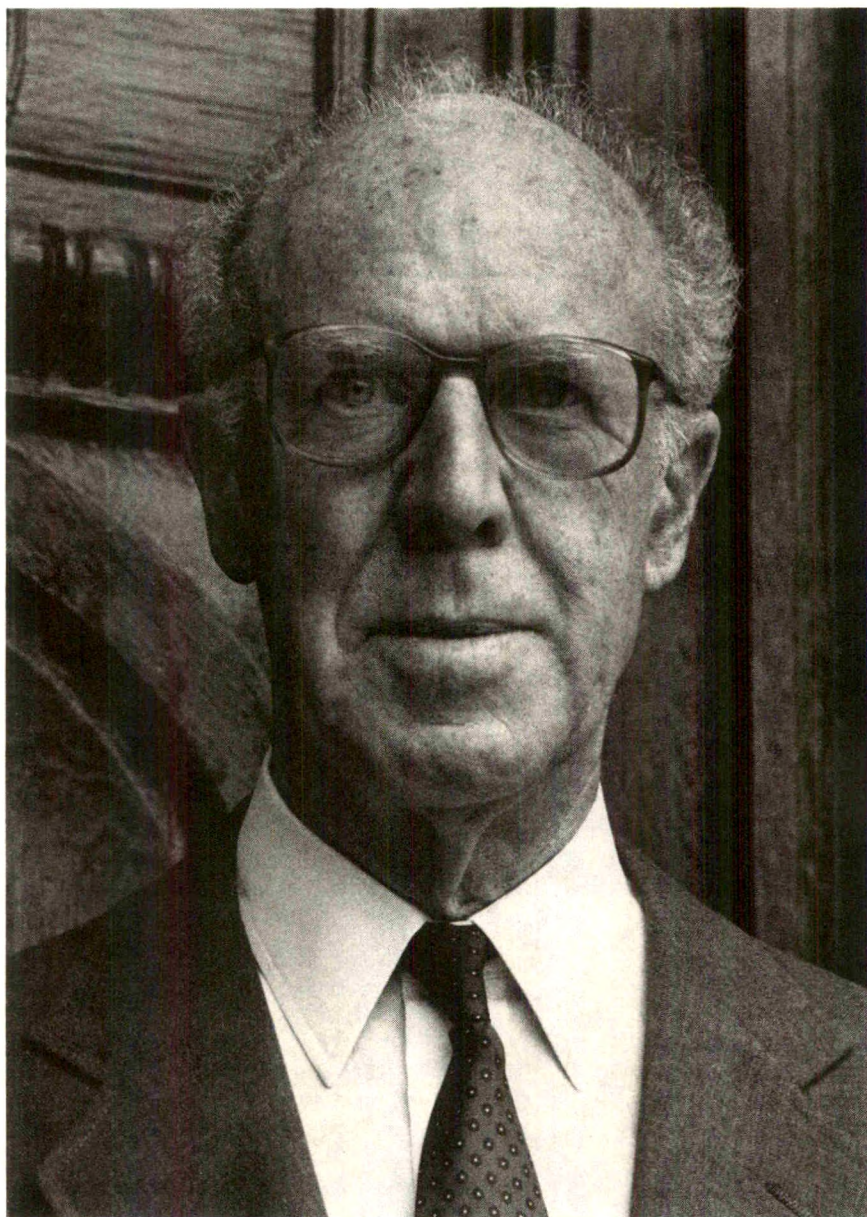
Orley Ashenfelter, Editor, *AER*, 209 Nassau Street, Princeton, NJ 08542-4607.

•Submission fee: \$50 for members; \$100 for nonmembers. Please pay with a check or money order payable in United States Dollars. Canadian and Foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Style guides will be provided upon request.

Editorial Statement

It is the policy of the *American Economic Review* to publish papers only where the data used in the analysis are clearly and precisely documented, are readily available to any researcher for purposes of replication, and where details of the computations sufficient to permit replication are provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary, or if, for some other reason, the above requirements cannot be met.

Number 92 of a series of photographs of past presidents of the Association



Gerard Debreu

The Mathematization of Economic Theory[†]

By GERARD DEBREU*

I

As the Second World War was drawing near its resolution, economic theory entered a phase of intensive mathematization that profoundly transformed our profession. In several of its main features that phase had no precedent, and it will have no successor. Assessing it requires a multidimensional analysis acknowledging the contributions to economics that were made, as well as the tensions among economists that were heightened.

The development of mathematical economics during the past half-century can be read in the total number of pages published each year by the leading periodicals in the field, an index that I will follow at first. From 1933, the date when they both started publication, to 1959, those periodicals were *Econometrica* and the *Review of Economic Studies*, and the index tells of the decline from a high point, above 700 pages in 1935 to the lowest point, below 400 pages in 1943–1944. But 1944 marked the beginning of a period of explosive growth in which *Econometrica* and the *Review of Economic Studies* were joined in 1960 by the *International Economic Review*, in 1969 by the *Journal of Economic Theory*, and in 1974 by the *Journal of Mathematical Economics*. In 1977, these five periodicals together published over 5,000 pages. During the period 1944–1977, the index more than doubled every nine years. By that measure, 1944 was a sharp turning point in the history of mathematical economics. It was also the year in which John von Neumann and Oskar

Morgenstern published the *Theory of Games and Economic Behavior*.

While the professional journals in the field of mathematical economics grew at an unsustainably rapid rate, the *American Economic Review* underwent a radical change in identity. In 1940, less than 3 percent of the refereed pages of its 30th volume ventured to include rudimentary mathematical expressions. Fifty years later, nearly 40 percent of the refereed pages of the 80th volume display mathematics of a more elaborate type.

At the same time, the mathematization of economists proceeded at an even faster pace in the 13 American departments of economics labeled by a recent assessment of research-doctorate programs in the United States (Lyle V. Jones et al., 1982) as “distinguished” or “strong” according to the scholarly quality of their faculties. Every year the Fellows of the Econometric Society (ES) certify new members by election into their international guild, which increased in size from 46 in 1940 to 422 in 1990. For those 13 departments together, the proportion of ES Fellows among professors was less than 1 percent in 1940; it is now close to 50 percent. It equals or exceeds 50 percent for six of them, which were among those assessed as the eight strongest. So mathematized a faculty expects its students to have what it considers to be minimal mathematical proficiency, and knowledge of calculus and linear algebra is required, or forcefully recommended, for admission to all 13 graduate programs.

Several scholarly recognitions lay additional emphasis on the role that mathematical culture is now playing in our profession. Of the 152 members of the economics section of the American Academy of Arts and Sciences, 87 are Fellows of the Econometric Society; and of the 40 members of the economics section of the National Academy of

[†]Presidential address delivered at the one-hundred third meeting of the American Economic Association, December 29, 1990, Washington, DC.

*Department of Economics, University of California at Berkeley, Berkeley, CA 94720.

Sciences of the United States, 34 are ES Fellows. From 1969 to 1990, 30 economics Nobel awards were made, and 25 of the laureates are, or were, ES Fellows. Since it was first presented to Paul Samuelson in 1947, the John Bates Clark medal of the American Economic Association has been given to 21 economists, of whom 20 are ES Fellows; and of the 26 living past presidents of our Association, 13 are ES Fellows.

One may wish that those counts had not been made. One may argue about points of their interpretation. But they belong in our common knowledge, and their thrust is unequivocal. They indicate how extensive the mathematization of economics and how deep the accompanying change of our field were over the past five decades.

The perception of the depth of that change is reinforced by a comparison of the levels of mathematics required in 1940 and in 1990 to follow the development of economic theory in every direction it was taking. Fifty years ago, basic undergraduate preparation in mathematics was almost always sufficient. Today, graduate training in mathematics is necessary. If, instead of being a follower, one wishes to be an active participant in that development along its most technical avenues, a high degree of mathematical professionalism is called for. Several faculty members of the 13 departments of economics mentioned previously were actually identified as mathematicians by their doctorates; four of them served as chairmen of those departments during the past 25 years. If still sharper focus brings out the intellectual leaders of that development, prominent among them is John von Neumann, one of the foremost mathematicians of his generation.

In that development process, mathematical economics was continuously redefined as new territories were included within its outward-moving frontier and as topics that were once at that frontier became standard parts of the graduate, if not of the undergraduate, economic-theory curriculum.

II

Before the contemporary period of the past five decades, theoretical physics had

been an inaccessible ideal toward which economic theory sometimes strove. During that period, this striving became a powerful stimulus in the mathematization of economic theory.

The great theories of physics cover an immense range of phenomena with a supreme economy of expression. Of this, James Clerk Maxwell (1865) had given a notable example, as he described the electromagnetic field by means of eight equations at the time when mathematical economics was born and came of age in the middle of the 19th century. This extreme conciseness is made possible by the privileged relationship that developed over several centuries between physics and mathematics. In turn, the former presented the latter with open problems, or found to questions raised by physical theory ready-made answers discovered by mathematicians in their abstract universe. Sometimes the causal linkage of research done in each one of the two fields could not easily be unraveled; and, on occasion, the same scientist made inextricably intertwined contributions to both disciplines.

The benefits of that special relationship were large for both fields; but physics did not completely surrender to the embrace of mathematics and to its inherent compulsion toward logical rigor. The experimental results and the factual observations that are at the basis of physics, and which provide a constant check on its theoretical constructions, occasionally led its bold reasonings to violate knowingly the canons of mathematical deduction.

In these directions, economic theory could not follow the role model offered by physical theory. Next to the most sumptuous scientific tool of physics, the Superconducting Super Collider whose construction cost is estimated to be on the order of $\$10^{10}$ (David P. Hamilton, 1990; see also *Science*, 5 October 1990), the experiments of economics look excessively frugal. Being denied a sufficiently secure experimental base, economic theory has to adhere to the rules of logical discourse and must renounce the facility of internal inconsistency. A deductive structure that tolerates a contradiction does so under the penalty of being useless,

since any statement can be derived flawlessly and immediately from that contradiction.

In its mathematical form, economic theory is open to an efficient scrutiny for logical errors. The rigor that has been reached as a consequence is in sharp contrast to the standards of reasoning that were accepted in the late 1930's. Few of the articles published then by *Econometrica* or by the *Review of Economic Studies* would pass the acid test of removing all their economic interpretations and letting their mathematical infrastructure stand on its own. The greater logical solidity of more recent analyses has contributed to the rapid contemporary construction of economic theory. It has enabled researchers to build on the work of their predecessors and to accelerate the cumulative process in which they are participating.

But a Grand Unified Theory will remain out of the reach of economics, which will keep appealing to a large collection of individual theories. Each one of them deals with a certain range of phenomena that it attempts to understand and to explain. When it acquires an axiomatic form, its explicit assumptions delimit its domain of applicability and make illegitimate overstepping of its boundary flagrant. Some of those theories take a comprehensive view of an economic system and bring insights into the solutions of several global problems. For instance, prices contribute to achieving an efficient use of resources, to equalizing supply and demand for commodities, and to preventing the formation of destabilizing coalitions. In every case, a theoretical explanation must be provided. The assumptions, which cannot be satisfied by all economic observations, are the present outcome of a continuing weakening process.

A global view of an economy that wants to take into account the large number of its commodities, the equally large number of its prices, the multitude of its agents, and their interactions requires a mathematical model. Economists have successfully constructed such a model because the central concept of the quantity of a commodity has a natural linear structure. The action of an agent can then be described by listing the quantity of its input or output for each

commodity (opposite signs differentiating inputs from outputs). That list can be treated as the list of the coordinates of a point in the linear commodity space. Similarly, the price system of an economy can be treated as a point in the linear price space, dual of the commodity space, whose dimension is also the number of commodities.

In those two linear spaces, the stage was set for sometimes dazzling mathematical developments that began with the elements of differential calculus and linear algebra and that gradually called on an ever broader array of powerful techniques and fundamental results offered by mathematics. Thus, the three roles of prices given earlier as instances were illuminated by basic mathematical theorems: the first, the achievement of an efficient use of resources, by results of convex analysis; the second, the equalization of supply and demand for commodities, by results of fixed point theory; the third, the prevention of the formation of destabilizing coalitions, by results of the theory of integration and of nonstandard analysis. In those three cases, the lag between the date of a mathematical discovery and the date of its application to economic theory decreased over time. It was notably short for nonstandard analysis, founded at the beginning of the 1960's by Abraham Robinson¹ and applied to economics by Donald Brown and Abraham Robinson (1972).

The last, and most recently developed, of those three instances can be chosen, as can either of the other two, for a more detailed illustration. Competition is perfect when every agent's influence on the outcome of economic activity is insignificant. The influence of their totality on that outcome is, however, significant. It is to solve the problem of aggregating negligible quantities so as to obtain a nonnegligible sum that integration was invented. In this perspective, the application of integration theory to the study of economic competition is entirely natural. That application requires the set of agents to be large—larger than the set of integers. Treating the set of the agents of an economy as the rich collection of the points

¹See the preface in Robinson (1966).

of an interval of real numbers has long been familiar in descriptions of economic data. It became familiar in economic theory as well after Robert J. Aumann (1964) showed that, in a pure exchange economy composed of insignificant agents, the formation of destabilizing coalitions is prevented if and only if all those agents base their decisions on a price system.

The concept of a convex set (i.e., a set containing the segment connecting any two of its points) had repeatedly been placed at the center of economic theory before 1964. It appeared in a new light with the introduction of integration theory in the study of economic competition: if one associates with every agent of an economy an arbitrary set in the commodity space and if one averages those individual sets over a collection of insignificant agents, then the resulting set is necessarily convex.² But explanations of the three functions of prices taken as examples can be made to rest on the convexity of sets derived by that averaging process. Convexity in the commodity space obtained by aggregation over a collection of insignificant agents is an insight that economic theory owes in its revealing clarity to integration theory.

An economist who experiences such an insight belongs to the group of applied mathematicians, whose values he espouses. Mathematics provides him with a language and a method that permit an effective study of economic systems of forbidding complexity; but it is a demanding master. It ceaselessly asks for weaker assumptions, for stronger conclusions, for greater generality. In taking a mathematical form, economic theory is driven to submit to those demands. The gains in generality that it has achieved as a result, in little more than a century, stand out when the first formulations of the theories of general equilibrium (Léon Walras, 1874–1877) and of the core of an economy (Francis Y. Edgeworth, 1881 pp. 34–8) are placed side by side with the recent treatments of those subjects to which *The New Palgrave* is an introduction and a

bibliographical key (John Eatwell et al., 1987–1989). Walras's consumers and producers have been freed from many of their constraining characteristics; Edgeworth's universe of two consumers and two commodities has been vastly expanded.

Mathematics also dictates the imperative of simplicity. It relentlessly searches for short transparent proofs and for the theoretical frameworks in which they will be inserted. Participating in that pursuit, economic theory was sometimes drawn by drives toward greater generality and toward greater simplicity in the same direction, rather than in opposite directions. Cohort after cohort, students of consumer theory have learned about the concept of decreasing marginal rate of substitution for two commodities on an indifference curve and about its extension to the multicommodity case. Notably more general, and notably simpler, is the concept of convexity of the set of points preferred to a given point in the commodity space. Welfare economics presents another instance. One of its main theorems formulates precisely the principle enunciated by Adam Smith (1776). If all the agents of an economy are in equilibrium relative to a price system, then they utilize their collective resources optimally. The proof of that theorem (Kenneth J. Arrow, 1951) has become so simple that it can be given without mathematical symbols. It is, at the same time, of utmost generality; in relating two basic concepts of economic theory to each other, it uses no assumption.

In its attempts to attain its many objectives, economic theory was helped by greater abstraction. Preference theory supplies an example again. Significant research efforts were expended on solutions of the integrability problem. That problem can be bypassed altogether, and greater simplicity can be achieved by moving from the commodity space to the more abstract space of the pairs of its points. In this space, whose dimension is twice the number of commodities, the pairs of commodity points indifferent to each other are now assumed to form a smooth (hyper)surface. As another instance of the generality permitted by abstraction, consider the notion of a commod-

²On this direct consequence of a theorem of A. A. Lyapunov, see Karl Vind (1964).

ity, which can be treated as a primitive concept, with an unspecified interpretation, in an axiomatic economic theory. A newly discovered interpretation can then increase considerably the range of applicability of the theory without requiring any change in its structure. Thus, by making the transfer of a good or service between two agents contingent on the state of the world that will obtain, Arrow (1953) made possible the immediate extension of the economic theory of certainty to an economic theory of uncertainty by a simple reinterpretation of the concept of a commodity. The theory of financial markets has been influenced by that view of uncertainty, and their practice has not been unaffected. Finally, take the problem of existence of a general equilibrium, once considered to be one of the most abstract questions of economic theory. The solutions that were proposed in the early 1950's paved the way for the algorithms for the computation of equilibria of Herbert E. Scarf (1973) and for several of the developments of applied general equilibrium analysis (Scarf and John B. Shoven, 1984). In this case, abstraction in economic theory led to the study of fundamental problems of great generality, but also to a broad range of applications.

III

The list of advances that the mathematization of economic theory helped or permitted is already long; and in one aspect it may appear lengthy. *Ceteris paribus*, one cannot prefer less to more rigor, lesser to greater generality, or complexity to simplicity; but other things are not equal, and in the estimate of many members of our Association the cost of that mathematization sometimes outweighs its benefit. Two of its presidential addresses notably confronted that difficult analysis and stressed the price that economics paid for its increased use of mathematics. Wassily Leontief's (1971) observations were factual, and Robert A. Gordon's (1976) comments relevant when they were made in 1970 and in 1975. They still are today, for, in spite of their authorities, enhanced by the platform from which they

were speaking, and in spite of the wide diffusion of their critiques, neither Leontief nor Gordon altered the course of the development they were assessing. In the past two decades, economic theory has been carried away further by a seemingly irresistible current that can be explained only partly by the intellectual successes of its mathematization.

Essential to an attempt at a fuller explanation are the values imprinted on an economist by his study of mathematics. When a theorist who has been so typed judges his scholarly work, those values do not play a silent role; they may play a decisive role. The very choice of the questions to which he tries to find answers is influenced by his mathematical background. Thus, the danger is ever present that the part of economics will become secondary, if not marginal, in that judgment.

The reward system of our profession reinforces the effects of that autocriticism. Decisions that shape the career of an economic theorist are made by his peers. Whether they are referees of a journal or of a research organization, members of an appointment or of a promotion committee, when they sit as judges in any capacity, their verdicts will not be independent of their own values. An economist who appears in their court rarely ignores his perception of those values. If he believes that they rate mathematical sophistication highly, and if he can prove that he is one of the sophisticates, the applause that he expects to receive will condition his performance.

The same effects are also amplified by the relentless pressure to publish exerted by his environment. There are indeed instances of extreme restraint in scientific publication, and some of them have become legend. The mathematical papers of Bernhard Riemann (1826–1866) take 506 pages in the volume that collected them (Riemann, 1876). The molecular structure of DNA was announced by James Watson and Francis Crick (1953) in a one-page article. But it is easier to explain those examples away than to follow them. The environment of a scholar demands papers, and the temptation to supply them without restraint may become over-

powering to an economic theorist who has developed proficiency in his research style. The precocious development of that proficiency is a comparative advantage that a mathematical approach bestows on him.

The spread of mathematized economic theory was helped even by its esoteric character. Since its messages cannot be deciphered by economists who do not have the proper key, their evaluation is entrusted to those who have access to the code. But acceptance of their technical expertise also implies acceptance of their values. Our profession may take pride in its exceptional intellectual diversity, one of whose clearest symbols is an Ely lecture given by an economic historian at a session chaired by a mathematical economist. Yet that diversity is strained by the increasing impenetrability to the overwhelming majority of our Association of the work done by its most mathematical members.

IV

The bond that ties economists together in their study of a common subject has not been tested only by differences in methodologies. It has also been tried by differences in ideologies. In their endeavors to make their field into a science, economists must renounce a favorite mode of thinking—wishful thinking; they must be impartial spectators of a play in which they are the actors. While they attempt to keep that inhuman stance, they are pressed to give immediate answers to societal questions of immense complexity and thereby to abandon the exacting slowness of the step-by-step scientific approach. Divisions according to methodologies and ideologies, criticism from outside and from inside, and intellectual fashions that sweep our discipline make each one of its steady developments remarkable. The mathematization of economic theory was one of them for a century and a half. During the past five decades it became one of the prime movers in the transformation of our field. The extent of that mathematization has given rise to discordant assessments of its effects and to attempts to change its heading. The quality of assess-

ments of the phase that economic theory underwent and the effectiveness of attempts to alter the course of its evolution will gain from a detailed analysis of the processes that led to its present state.

REFERENCES

- Arrow, Kenneth J., "An Extension of the Basic Theorems of Classical Welfare Economics," in J. Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, 1951, 507–32.
- , "Le Rôle des Valeurs Boursières pour la Répartition la Meilleure des Risques," in *Econométrie*, Paris: Centre National de la Recherche Scientifique, 1953, 41–8.
- Aumann, Robert J., "Markets with a Continuum of Traders," *Econometrica*, January–April 1964, 32, 39–50.
- Brown, Donald and Robinson, Abraham, "A Limit Theorem on the Cores of Large Standard Exchange Economies," *Proceedings of the National Academy of Sciences USA*, May 1972, 69, 1258–60.
- Eatwell, John, Milgate, Murray and Newman, Peter, eds., *General Equilibrium*, New York: Macmillan, 1987–1989.
- Edgeworth, Francis Y., *Mathematical Psychics*, London: Paul Kegan, 1881.
- Gordon, Robert A., "Rigor and Relevance in a Changing Institutional Setting," *American Economic Review*, March 1976, 66, 1–14.
- Hamilton, David P., "The SSC Takes On a Life of Its Own," *Science*, 17 August 1990, 249, 731–2.
- Jones, Lyle V., Lindzey, Gardner and Coggeshall, Porter E., eds., *An Assessment of Research-Doctorate Programs in the United States: Social and Behavioral Sciences*, Washington DC: National Academy of Sciences Press, 1982.
- Leontief, Wassily, "Theoretical Assumptions and Non-observed Facts," *American Economic Review*, March 1971, 61, 1–7.
- Maxwell, James C., "A Dynamical Theory of the Electromagnetic Field," *Philosophical Transactions of the Royal Society of Lon-*

- don*, 1865, 155, 459–512.
- Riemann, Bernhard**, *Gesammelte Mathematische Werke und Wissenschaftlichen Nachlass*, Leipzig: Teubner, 1876.
- Robinson, Abraham**, *Non-Standard Analysis*, Amsterdam: North-Holland, 1966.
- Scarf, Herbert E.**, (with the collaboration of T. Hansen), *The Computation of Economic Equilibria*, New Haven: Yale University Press, 1973.
- _____ and **Shoven, John B.**, *Applied General Equilibrium Analysis*, New York: Cambridge University Press, 1984.
- Smith, Adam**, *An Inquiry into the Nature and Causes of the Wealth of Nations* (2 volumes), London: W. P. Strahan and T. Cadell, 1776.
- Vind, Karl**, "Edgeworth Allocations in an Exchange Economy with Many Traders," *International Economic Review*, May 1964, 5, 165–77.
- von Neumann, John and Morgenstern, Oskar**, *Theory of Games and Economic Behavior*, Princeton: Princeton University Press, 1944.
- Walras, Léon**, *Eléments d'Economie Politique Pure*, Lausanne: L. Corbaz, 1874–1877.
- Watson, James D. and Crick, Francis H. C.**, "A Structure for Deoxyribose Nucleic Acid," *Nature*, 25 April 1953, 171, 737–8.
- Science*, 5 October 1990, 250, 28.

An Econometric Analysis of U.K. Money Demand in *Monetary Trends in the United States and the United Kingdom*

by Milton Friedman and Anna J. Schwartz

By DAVID F. HENDRY AND NEIL R. ERICSSON *

This paper evaluates an empirical model of U.K. money demand developed by Milton Friedman and Anna J. Schwartz in Monetary Trends in the United States and the United Kingdom. Testing reveals misspecification and hence the potential for an improved model. Using recursive procedures on their annual data, we obtain a better-fitting, constant, dynamic error-correction (cointegration) model. Results on exogeneity and encompassing imply that our money-demand model is interpretable as a model of money but not of prices, since its constancy holds only conditionally on contemporaneous prices. (JEL 210)

In their 1982 book, *Monetary Trends in the United States and the United Kingdom: Their Relation to Income, Prices, and Interest Rates, 1867–1975*, Milton Friedman and Anna Schwartz present a wealth of empiri-

cal findings as support for a range of economics hypotheses.¹ To isolate the longer-term relationships that are their primary concern, Friedman and Schwartz estimate their empirical models from data transformed by averaging annual observations over phases of NBER reference cycles. In our analysis, we first replicate one of their preferred models for U.K. money demand on the phase-average data and then evaluate it econometrically to investigate aspects of model specification about which the data are most informative. The outcome indicates the potential for an improved equation but does not entail the form of respecification required. In seeking to construct an improved model, we formulate an equation that integrates long-run properties with short-run dynamics, based on the recent merging of the theories of error-correction and cointegration. Moreover, because phase-averaging may entail a loss of information about relevant parameters, we return to analyzing the underlying annual observations. Finally, the resulting model is critically evaluated.

*Hendry is Professor of Economics at Nuffield College, Oxford, England OX1 1NF, and Visiting Research Professor at the Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27706; Ericsson is a staff economist in the International Finance Division, Federal Reserve Board, Washington, DC 20551. This research was supported in part by U.K. Economic and Social Research Council (E.S.R.C.) grants HR8789 and B00220012. We are grateful for the financial assistance from the E.S.R.C. although the views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting those of the E.S.R.C., the Board of Governors of the Federal Reserve System, or other members of their staffs. Helpful discussions with and comments from Chris Allsopp, Julia Campos, Nicholas Dimsdale, Rob Engle, John Flemming, Milton Friedman, David Germany, John Geweke, Dale Henderson, Kevin Hoover, Jeroen Kremers, Bonnie Loopesko, Mary Morgan, John Muellbauer, Adrian Neale, Charles Nelson, Jean-François Richard, Ken Rogoff, John Taylor, Ken Wallis, Arnold Zellner, and several anonymous referees are gratefully acknowledged. We are indebted to Jeroen Kremers, Cathy Fitzgerald, and Aileen Liu for their invaluable assistance with the calculations. This is an extensively revised and shortened version of papers presented to the Bank of England's Panel of Academic Consultants (Hendry and Ericsson, 1983) and circulated by the Federal Reserve Board (Hendry and Ericsson, 1985).

¹For helpful reviews and summaries of Friedman and Schwartz (1982), see inter alia Charles Goodhart (1982), Thomas Mayer (1982), Tim Congdon (1983), Basil Moore (1983), and Michael Artis (1984).

The sequence in the previous paragraph involves a natural progression from model discovery to model evaluation through replication and testing, and then via new conjectures back to discovery, seeking models that account for previous findings and explain additional phenomena. The paper's main objective is to achieve this last goal for U.K. money-demand models over the period 1878–1970. An additional objective is to exposit an econometric framework that makes precise the notion of an improved model, explains the construct of accounting for previous findings (denoted encompassing), and delineates the criteria for model evaluation. This enables us to clarify the concepts of exogeneity, invariance, encompassing, and cointegration in the context of an important economics debate. The empirical model reported below improves upon the phase-average U.K. money-demand equation from Friedman and Schwartz, our first annual-data specification (proposed in 1983), and those in the studies that the latter stimulated, thus illustrating a progressive research methodology in action.

In Section I, we consider the data and data transformations used by Friedman and Schwartz (1982). In Section II, we discuss the money-demand relationships they estimated from the phase-average data for the United Kingdom and evaluate many of their empirical claims about the selected money-demand equation. The statistical framework and its associated concepts are briefly explicated in Section III to clarify the joint destructive and constructive roles for tests. In Section IV, we constructively apply that approach to modeling money demand using the annual U.K. data series. Our conclusions are presented in Section V.

I. The Data Series and Transformations

In Friedman and Schwartz's (1982) and our studies of U.K. money demand, the basic data series are annual values for the United Kingdom from 1871 to 1975 of the broad money stock (M), real net national income (I), the price level (P), short-term and long-term nominal interest rates (RS and RL), population (N), and high-powered

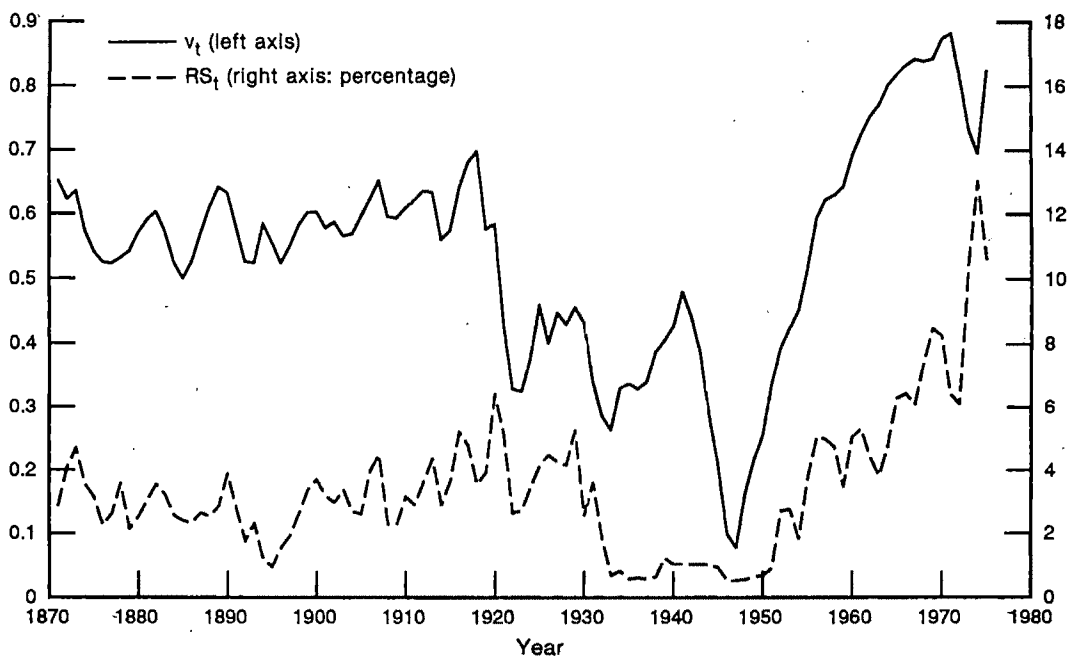


FIGURE 1. THE LOG OF THE VELOCITY OF MONEY (v_t) AND THE YIELD ON THREE-MONTH BILLS (RS_t) (ANNUAL DATA)

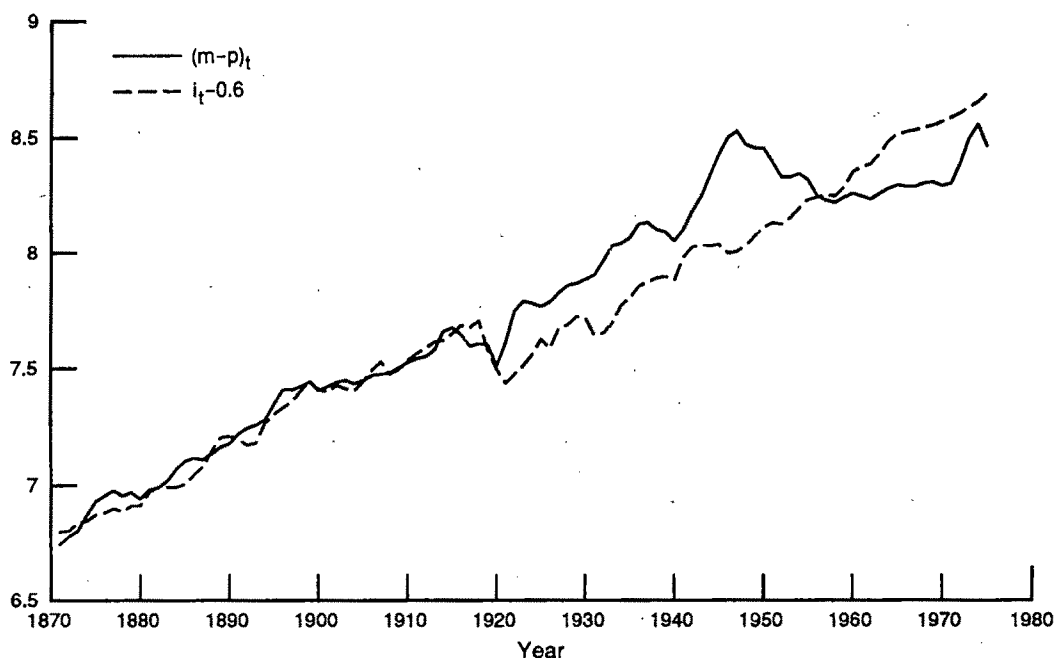


FIGURE 2. THE LOG OF THE REAL MONEY STOCK $([m - p]_t)$ AND THE LOG OF REAL INCOME (i_t) (ANNUAL DATA)

money (H); and the price level (P^*) in the United States. All series are in Friedman and Schwartz's tables 4.8 and 4.9; for details of construction and definition of the series, see both Friedman and Schwartz (1982 Ch. 4, 5) and our Appendix, which note reservations on the measurement and interpretation of those series. Unless otherwise indicated, capital letters denote both the generic name and the level; logs of scalars are denoted by lowercase letters. Figures 1, 2, and 3 respectively show the annual time series for velocity V ($\equiv PI/M$) and RS, real money (M/P) and real income, and nominal money and prices.²

From Figure 1, three distinct episodes are discernible in the behavior of velocity, coinciding with historical periods before the First World War (1871–1914), interwar (1918–1939), and after the Second World War (1945–1975). Initially, velocity and RS cycle around constant levels. Then, in the inter-

war period, velocity is lower and more volatile. It falls during the Second World War, reaching its lowest observed level in 1947, corresponding to the lowest level for interest rates (of 0.5 percent per annum). Next, velocity trends upwards, more than doubling by 1970, while interest rates fluctuate increasingly around an upward trend. Finally, between 1971 and 1975 velocity falls while RS rises.

Some insights into this behavior of velocity can be gleaned from Figure 2, which plots two of its constituents, $m - p$ and i . The interwar period begins with a sharp fall in i being matched by a large rise in $m - p$, which then remains systematically above i with its greatest departure from i in 1947. Thereafter, $m - p$ falls steadily until around 1960, whereas i rises considerably, and then $m - p$ "levels off" until its sharp rise in 1971. Thus, between 1947 and 1970 measured velocity more than doubles because $m - p$ drops while i rises. Lastly, Figure 3 shows nominal money and the price level. These series are highly correlated in levels although, overall, money increases 5.6 times

²When there is no loss of clarity, we often refer to v ($= p + i - m$) as velocity.

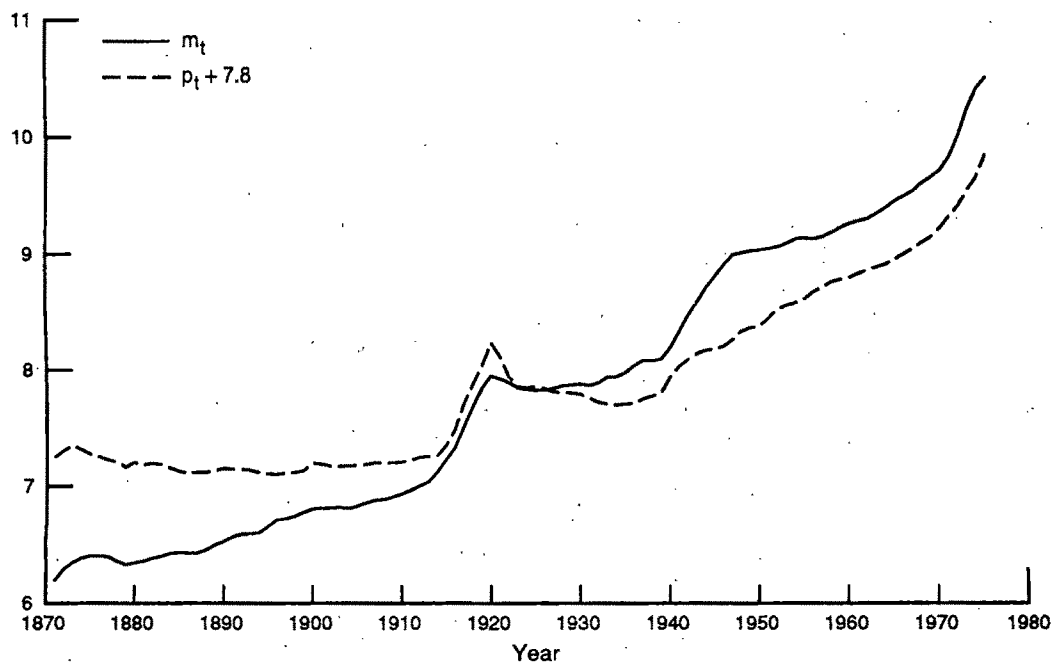


FIGURE 3. THE LOG OF THE NOMINAL MONEY STOCK (m_t) AND THE LOG OF THE IMPLICIT PRICE DEFLATOR (p_t) (ANNUAL DATA)

relative to p (real income increases by 6.7 times), and during 1920–1922 a large fall in p occurs alongside a small fall in m . To represent adequately the century of data, an econometric model of U.K. money demand must encompass such features.

Since Friedman and Schwartz aim to examine the long-period movements in the data, they do not analyze these annual series directly but first transform them by averaging separately over contraction and expansion phases of data-selected choices of reference business cycles. From some hundred years of annual U.K. observations, Friedman and Schwartz identify 37 such phases with average lengths of 2.1 and 3.4 years respectively for contraction and expansion phases. These phase averages, weighted by a function of duration, are the basic units of analysis (pp. 75–9 and table 5.8)³ and are analyzed both in levels and as

rates of change. The latter are calculated from least-squares slopes fitted to triplets of overlapping phases (weighted appropriately) and are arithmetically nearly the same as differences from one expansion (or contraction) to the next.⁴ The data to which these moving-average filters are applied are the logarithms of money, prices, incomes, and population, and the original values of interest rates. We denote phase-average data by an overbar (e.g., \bar{m}_j for the j th of J phase-average observations of the logarithm of money stock). These data are the basis for the analysis in Section II, whereas in Section IV we utilize the original annual time series. The relationship between the phase-average and the annual series for velocity is shown in Figure 6, discussed below.

³Unreferenced page numbers and table numbers when “table” is uncapitalized refer to Friedman and Schwartz (1982).

⁴There are typographical errors in the reported values for “rates of change” in table 5.10 affecting observations 18–21. We corrected those errors following the procedures described by Friedman and Schwartz (1982 section 3.2); our amendments are reported in Hendry and Ericsson (1985 appendix F).

II. Testing the U.K. Money-Demand Model on the Phase-Average Data

In this section, we summarize the central empirical results obtained by Friedman and Schwartz (1982) for U.K. money demand using the phase-average data, replicate one of their preferred models, and econometrically evaluate it. Table 1 reports estimates of four equations out of those that Friedman and Schwartz record for the United Kingdom as establishing evidence concerning the demand for money, and the influences of money on income and on prices. "Levels" and "rates of change" equations are analyzed by Friedman and Schwartz but, as they see little to choose between them (e.g., see p. 286), we focus on the equations in levels: full details can be found in Friedman and Schwartz. We summarize their interpretation of the estimates in Table 1 as follows. Equation (a) is their preferred U.K. money-demand equation and has a nearly

unit income elasticity (pp. 280–4). From equation (d), I does not depend upon M , whereas P does [eq. (c)] and, hence, so does PI [eq. (b)] (cf. pp. 422, 351).

Equations (a)–(d) represent types of models (of money, nominal incomes, prices, and output, respectively), but from our perspective their joint existence raises difficulties. Because M is a regressor in equations (b)–(d), M must be (weakly) exogenous if the coefficients in these regressions are to be interpreted as parameters of interest (cf. Robert Engle, Hendry, and Jean-Francois Richard, 1983). With M exogenous, equation (a) is determining P , not M ; but so is equation (c). Conversely, if M does not maintain its exogeneity status in equation (a), P and M are jointly determined, and hence equation (c) must be invalid. However, since Friedman and Schwartz take (a) as one of their "final equations" for U.K. money demand (p. 281), we will evaluate it on that basis.

TABLE 1—EQUATIONS FOR U.K. MONEY DEMAND, NOMINAL INCOMES, PRICES, AND OUTPUT FROM FRIEDMAN AND SCHWARTZ (1982) (PHASE-AVERAGE DATA)

Equation	Source
(a) $(\bar{m} - \bar{p} - \bar{n}) = 0.16 + 0.88(\bar{i} - \bar{n}) - 11.16\bar{RN} - 0.22G(\bar{p} + \bar{i}) + 1.4\bar{W} + 21\bar{S}$ (0.08)(18.13) (3.42) (0.74) (2.38) (7.56) $(\hat{\sigma} = 5.54, R^2 = 0.970, dw = 1.51, \eta_1[18, 12] = 6.3)$	p. 282
(b) $(\bar{p} + \bar{i}) = 0.38 + 1.01\bar{m} + 14.17\bar{RN} + 0.53G(\bar{p} + \bar{i}) - 1.1\bar{W} - 19\bar{S}$ $(\hat{\sigma} = 5.99, R^2 = 0.9977, dw = 1.33, \eta_1[18, 12] = 5.7)$	p. 349
(c) $\hat{p} = -5.99 - 0.015\bar{i} + 1.02\bar{m} + 0.94\bar{RS} - 1.10G(\bar{p} + \bar{i})$ (31.1) (7.4) (17.4) (0.9) (2.8) $(\hat{\sigma} = 6.0, dw = 1.01, \eta_1[15, 10] = 3.1)$	p. 420
(d) $\hat{i} = 6.50 + 0.017\bar{i} - 0.05\bar{m} + 2.34\bar{RS} + 2.20G(\bar{p} + \bar{i})$ (34.2) (8.3) (0.9) (2.4) (5.7) $(\hat{\sigma} = 5.9, dw = 0.97, \eta_1[15, 10] = 2.1)$	p. 420

Notes: Notation is as in Sections I and II, but here parentheses enclose t values, as reported in the original text. No t ratios or standard errors appear in Friedman and Schwartz (1982) for the equation labeled (b) above. $RN = (RS) \cdot H/M$; $G(\cdot)$ denotes a rate of change; and \bar{W} and \bar{S} are the dummies for postwar adjustment and demand shift, rescaled by 1/100 (see Appendix). Phase averages 12–14 and 24–26 are omitted in (c) and (d). Here and in Table 3, values of $\hat{\sigma}$ with logarithmic regressands are quoted as approximate percentages relative to the level of the regressand in its original units [e.g., if $\log(Y)$ is the regressand, $\hat{\sigma}$ is a percentage of Y]. Values for dw and η_1 are our calculations and are not reported in Friedman and Schwartz. Our equations replicating (c) and (d) had somewhat smaller values of $\hat{\sigma}$ than those in Friedman and Schwartz.

We could closely replicate equation (a):

$$\begin{aligned}
 (1) \quad & \overline{(\bar{m} - \bar{p} - \bar{n})}_j \\
 & = 0.012 + 0.885(\bar{i} - \bar{n})_j \\
 & \quad (0.19) \quad (0.049) \\
 & - 11.21\overline{RN}_j - 0.22G(\bar{p} + \bar{i})_j \\
 & \quad (3.3) \quad (0.29) \\
 & + 1.37\overline{W}_j + 20.6\overline{S}_j \\
 & \quad (0.58) \quad (2.7)
 \end{aligned}$$

($J = 36$, $\hat{\sigma} = 5.66$ percent, $dw = 1.51$, $R^2 = 0.97$) where \overline{RN} is the differential between \overline{RS} and the "own-yield" on money, $G(\bar{p} + \bar{i})$ is the growth rate of nominal income (interpreted by Friedman and Schwartz as proxying for the rate of return on physical assets), \overline{W} is a dummy for "postwar adjustment" (p. 228), and \overline{S} is a data-based dummy for "[a]n upward demand shift, produced by economic depression and war..." (p. 281) during 1921–1955. \overline{S} captures a 21-percent shift in (1) and so accounts for much of the variation in real per capita money, conditional on real per capita income. Figure 4 shows the fitted and actual values for $(\bar{m} - \bar{p} - \bar{n})$ derived from (1). Throughout, estimated standard errors are shown in parentheses (except in Table 1), $\hat{\sigma}$ is the standard deviation of the residuals [e.g., in (1), as a percentage of $\overline{M}/(\overline{PN})$], adjusted for degrees of freedom, and R^2 is the squared multiple correlation coefficient for the corresponding unweighted regression.⁵

Having replicated equation (a) by (1), we evaluate (1) against a range of alternative models. A summary of our approach is provided in Section III; here, we focus on the usual concerns of parameter constancy, price homogeneity, omitted variables, homoscedasticity, and normality. Once an empirical model has been formalized, then,

conditional on treating it as provisionally valid, many empirical phenomena are excluded. Consequently, a model is testable against the potential occurrence of such phenomena, *using tests that would be valid given the assumptions of that model*. The resulting tests have central distributions with approximately 5-percent critical levels under the null hypothesis of correct specification and noncentral distributions under their respective alternatives. Additionally, they may have power to reject alternatives other than those against which they were designed or have larger implicit than explicit null hypotheses (cf. Grayham Mizon and Richard, 1986).

Constancy is a major issue for money-demand equations (see John Judd and John Scadding, 1982), and Friedman and Schwartz regard their own U.K. money-demand equation as being constant: "[A] more sophisticated analysis [than the simple quantity theory] reveals the existence of a stable demand function for money covering the whole of the period we examine" (p. 624; see also pp. 7, 14, 64, 283). However, they do not formally *test* for constancy, and many investigators would regard the need for the data-based shift dummy \overline{S} spanning one-third of the sample as *prima facie* evidence against the model's constancy. Refitting (1) to the first half of the sample and predicting over the second half (doing each with *both* \overline{S} and \overline{W}), we tested for constancy using Gregory Chow's (1960 pp. 594–5) statistic: that yielded $\eta_1[18, 12] = 6.3$, which exceeds the 1-percent point of the F distribution. Although the one-step-ahead 95-percent confidence interval based on $\hat{\sigma}$ is larger than ± 11 percent, parameter nonconstancy can be detected, revealing an aspect about which the present data set is informative. The values of $\hat{\sigma}$ for the corresponding subperiods are markedly different: 2.8 percent and 6.0 percent, respectively. Even if these differences were due solely to heteroscedasticity, the inferences that Friedman and Schwartz draw from their regressions would be invalid because the t ratios are biased. Moreover, nonconstancy is not restricted to the residual variance. Figure 5 records the recursive estimates of

⁵Like those in Friedman and Schwartz, all our estimates using phase-average data are based on weighted least squares, correcting for the different phase lengths. However, parameter estimates are not very different whether ordinary or weighted least squares are used. All empirical calculations are from PC-GIVE (see Hendry, 1989).

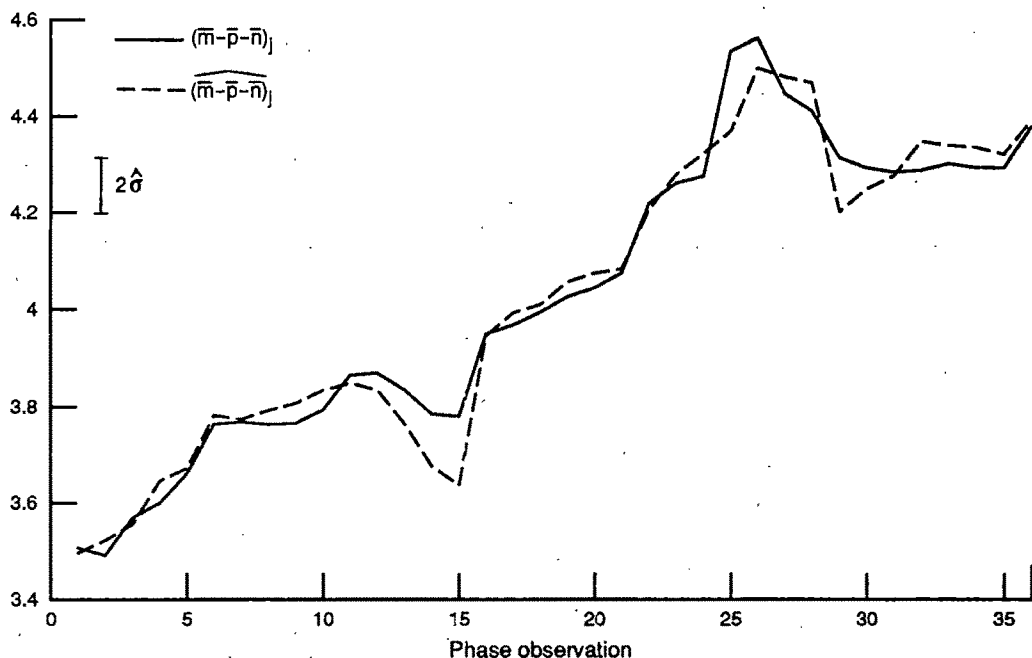


FIGURE 4. EQUATION (1): ACTUAL AND FITTED VALUES FOR $(\bar{m} - \bar{p} - \bar{n})_j$ (THE TRANSFORMED REGRESSAND)

the interest rate coefficient from phases 16–36, together with a confidence region based upon plus-or-minus twice its estimated standard error at each sample size.⁶ As the sample size increases, the estimated coefficient changes substantially relative to its estimated uncertainty, with the final estimate lying outside the initial confidence interval. Such evidence refutes constancy in Friedman and Schwartz's reported model.

Since the dependent variable in (1) is $\bar{m} - \bar{p} - \bar{n}$, price homogeneity can be tested under two distinct maintained hypotheses about exogeneity, namely, whether M or P

is exogenous. Treating M as endogenous and relaxing unit price homogeneity in (1), we obtain

$$\begin{aligned}
 (2) \quad & \widehat{(\bar{m} - \bar{p} - \bar{n})}_j \\
 &= 1.01 + 0.68(\bar{i} - \bar{n})_j + 0.128\bar{p}_j \\
 &\quad (0.38) \quad (0.08) \quad (0.043) \\
 &\quad - 18.4\overline{RN}_j - 0.04G(\bar{p} + \bar{i})_j \\
 &\quad (3.8) \quad (0.26) \\
 &\quad + 1.05\overline{W}_j + 16.3\overline{S}_j \\
 &\quad (0.53) \quad (2.8)
 \end{aligned}$$

⁶In textbook notation, Figure 5 plots the i th coefficient and its estimated standard error: $\hat{\beta}_{ij} \pm 2\text{ESE}(\hat{\beta}_{ij})$, $j = 16, \dots, 36$, where $\hat{\beta}_j = (\mathbf{X}_j' \mathbf{X}_j)^{-1} \mathbf{X}_j' \mathbf{Y}_j$, $\mathbf{X}_j = (\mathbf{x}_1 \dots \mathbf{x}_j)'$, $\mathbf{Y}_j = (y_1 \dots y_j)'$, and $\text{ESE}(\hat{\beta}_{ij}) = \sqrt{\text{diag}_i[\hat{\sigma}_j^2 (\mathbf{X}_j' \mathbf{X}_j)^{-1}]}$. Under the null hypothesis of correct specification, $\hat{\beta}_{ij} \rightarrow \beta_i$ and $\text{ESE}(\hat{\beta}_{ij}) \rightarrow 0$ as $j \rightarrow \infty$ (see Andrew Harvey [1981 pp. 54–9] on the calculation of recursive least squares and associated test statistics and Kerry Patterson [1986] for an application). For annual data, the index j becomes t .

($J = 36$, $\hat{\sigma} = 5.03$ percent, $\text{dw} = 1.85$). The coefficient on \bar{p}_j is significant at the 99-percent level. If, instead, the exogeneity of M is maintained so that (1) purports to explain \bar{p}_j , then adding \bar{m}_j and \bar{n}_j to (1) to relax the homogeneity assumption yields $F_{[2,28]} = 20.56$, again rejecting the specification of (1).

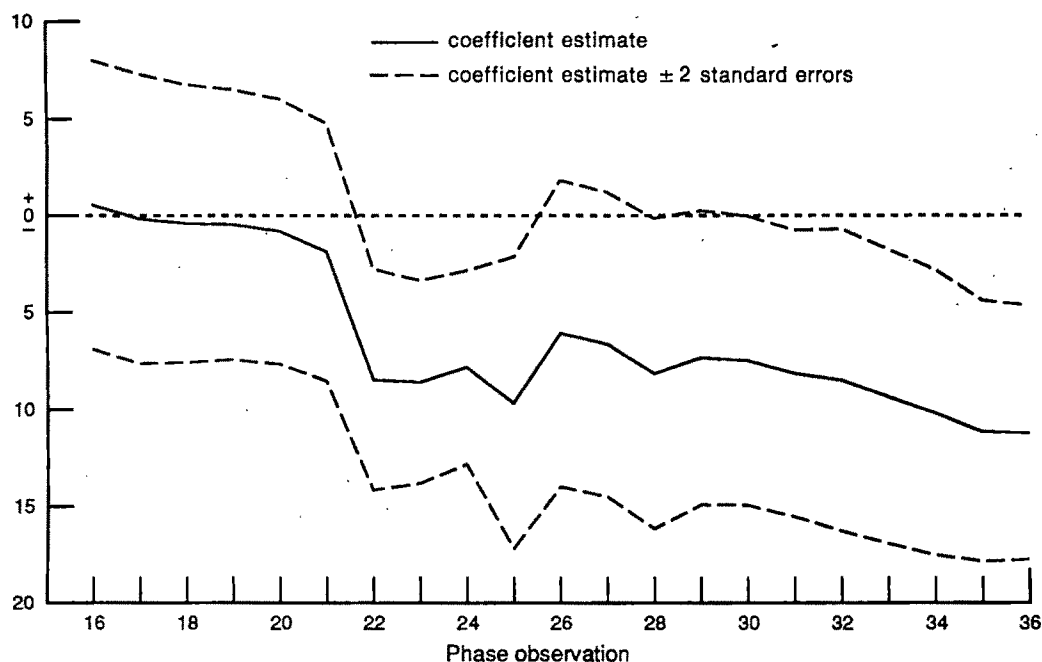


FIGURE 5. EQUATION (1): RECURSIVE ESTIMATION OF THE COEFFICIENT ON \overline{RN}_j AND ITS ESTIMATED STANDARD ERROR

Next, to test for the absence of a trend, we estimate (3):

$$\begin{aligned}
 (3) \quad & \overline{(\bar{m} - \bar{p} - \bar{n})}_j \\
 & = -13.92 + 0.14(\bar{i} - \bar{n})_j \\
 & \quad (2.99) \quad (0.16) \\
 & -16.0\overline{RN}_j + 0.23G(\bar{p} + \bar{i})_j \\
 & \quad (2.7) \quad (0.24) \\
 & +1.80\overline{W}_j + 6.1\overline{S}_j + 0.0090\bar{i}_j \\
 & \quad (0.46) \quad (3.7) \quad (0.0019)
 \end{aligned}$$

($J = 36$, $\hat{\sigma} = 4.35$ percent, $dw = 1.99$). The coefficient on \bar{i}_j is statistically significant. By way of comparison, when Friedman and Schwartz tested for trends and price homogeneity in more restrictive models, they did not obtain rejections on this data set (pp. 253–9).

Although Friedman and Schwartz did not test for it, normality of the disturbances is also rejected: Carlos Jarque and Anil Bera's (1980) $\chi^2_{[2]}$ statistic for testing normality via

residual skewness and excess kurtosis is $\xi_3[2] = 6.9$ in (1).⁷ Testing the normality of the disturbances is of interest for two reasons. First, Friedman and Schwartz often base their statistical inferences on critical values of the F distribution (e.g., see pp. 232, 236), and given the small sample size ($J = 36$), those inferences may be affected substantively by nonnormal disturbances. Second, the Jarque-Bera statistic also has power against other alternatives, as suggested by it being insignificant when unit homogeneity of money with respect to prices is not imposed.

Taking this evidence together, equation (1) is not an adequate characterization of the data and is not consistent with the hypothesis of a constant money-demand equa-

⁷Given that the number of regressors fitted is often large relative to the sample size, we modify Jarque and Bera's (1980 p. 257) statistic to be $\xi_3[2] = [(J - k)/6] \cdot [\widehat{SK}^2 + \widehat{EK}^2/4]$, where SK and EK are skewness and excess kurtosis, respectively. $\xi_3[2]$ is asymptotically distributed as $\chi^2_{[2]}$ under the null hypothesis of normality, in which case $SK = EK = 0$.

tion, homogeneous of degree one in prices over the last century, even though Friedman and Schwartz state that "[t]his parallelism⁸ is a manifestation of the stable demand curve for money plus the excellence of the simple quantity theory approximation" (p. 7). The equations for nominal income and prices in Friedman and Schwartz were not further investigated because, as explained in Hendry and Ericsson (1985 appendix D), they are approximately renormalizations of their money-demand regression and so do not provide additional inferences (cf. pp. 344, 417). Additionally, any inferences based on (b)–(d) are hazardous because the split-sample Chow statistics for (b) and (c) are highly significant, and the Durbin-Watson statistics in (c) and (d) are less than the 5-percent lower bound.

When interpreting the outcomes of the above tests, four points should be borne in mind. First, the rejections are not mutually independent sources of information, since the implicit assumptions of any given test are shown to be invalid by the outcomes on other tests. Secondly, *none of the relevant hypotheses could have been tested by Friedman and Schwartz for this equation without their having obtained a rejection*. This shows the power of statistical evaluation to reveal the potential for model improvement and the ability of the data to discriminate between empirical models. Thirdly, discovering that their empirical model has a variety of specification problems has no implications for the existence (or otherwise) of a correctly specified empirical model of money demand being homogeneous in prices and having constant parameters. Finally, rejection of any null hypothesis does not imply that the alternative is correct. For example, constancy may have been rejected because of dynamic misspecification, and not because the underlying money-demand model has nonconstant parameters; or, in small samples, an F test could yield a misleading

outcome because of a highly nonnormal error distribution. These problems frequently arise in approaches that proceed by generalizing simple models (see Hendry, 1979).

Thus, we disagree with the methodology adopted by Friedman and Schwartz of basing many of their inferences on the analysis of simple empirical models:

Our ultimate objective is an explanation of the behavior of velocity, which is to say, of the quantity of money demanded, that takes account simultaneously of all the variables affecting velocity. Nonetheless, we believe that this ultimate objective is better approached indirectly, by examining variables one or two at a time, than by what has become the prevailing fashion in econometric work, the immediate computation of multiple regressions including all variables that can reasonably be regarded as relevant. We believe that the indirect approach yields insights that cannot be obtained from the more sweeping approach—that multiple correlations with many variables are almost impossible to interpret correctly unless they are backed by more intensive investigations of smaller sets of variables. Accordingly, we shall proceed in the following sections to consider variables one or two at a time, and reserve to section 6.7 estimating their simultaneous effect. [footnote:] The indirect approach played a critical role in the formulation of the multiple regressions that we calculate in section 6.7. . . . (p. 215)

This quote raises four methodological issues. First, the claim that simple models yield insights is frequently mistaken, since the associated measures of relationship and of uncertainty are misleading unless such models are coherent with the data. Secondly, it is incorrect to test hypotheses in one model and infer that the outcome will hold in generalizations of that model unless all the additional influences are orthogonal to those already included and remain so throughout the sample. Thirdly, although we concur that it is difficult to interpret

⁸Of nominal income and the nominal quantity of money, and of the rate of change of nominal income and the rate of change of the nominal quantity of money.

multiple regressions, this does not justify fitting misspecified submodels. Finally, the penultimate sentence of the above quote conflates the issues of estimation and modeling by seeming to imply that the only step remaining after examining the smaller sets is to estimate their joint effect: that would be true only if there were no interactions. The crucial issue is that a reject outcome at any stage of modeling invalidates *all previous inferences*, so that decision-taking during model generalizations is ill-founded. In Section III, we briefly describe an alternative methodology which resolves many of these issues and also reveals that there are other approaches to understanding general models. Moreover, a constructive empirical task remains to be undertaken, and we turn to that in Section IV, using the model developed by Friedman and Schwartz as a benchmark against which to compare other equations.

Before proceeding, we address the question of whether to use the annual observations or the phase-average data, the latter

being those which Friedman and Schwartz chose to analyze.

In order to isolate the longer-term relations that are our primary concern, we have converted our basic data, which are annual, into a form designed to be free from the shorter-term movements that are called business cycles. ... The device we have used to free the data from cyclical fluctuations is to take as our basic observation an average of annual observations over a cycle phase... (p. 13)

This device is taken from the NBER approach to business-cycle analysis, as documented by Arthur Burns and Wesley Mitchell (1946). Friedman and Schwartz argue that phase-averaging reduces serial correlation arising from the business cycle (p. 78) and attenuates measurement errors (p. 86). Doing so is important to sustain the validity of their statistical analyses, since, for example, no explicit account is made for biases in coefficient estimates and estimated

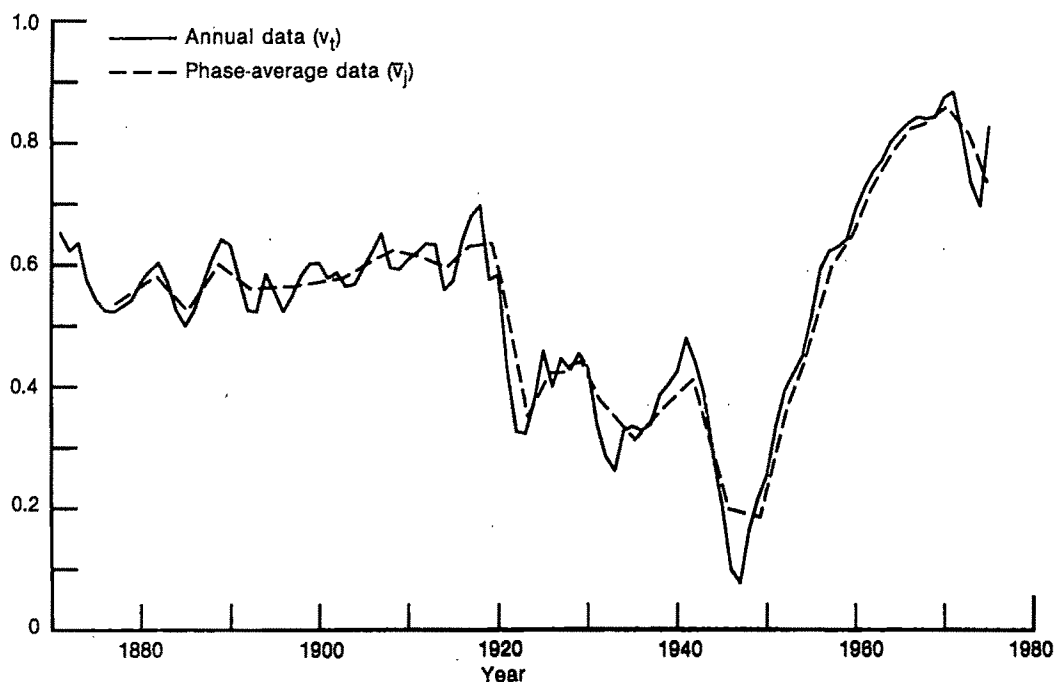


FIGURE 6. THE LOG OF THE VELOCITY OF MONEY: ANNUAL AND PHASE-AVERAGE DATA

standard errors resulting from residual autocorrelation. Friedman and Schwartz claim that phase-averaging successfully isolates longer-run behavior: "All in all, we conclude that our phase bases do eliminate the bulk of the systematic cyclical fluctuation" (p. 81).

The use of phase-average data raises three distinct issues: the theoretical statistical effects of phase-averaging (qua aggregation), the observed effects of phase-averaging on Friedman and Schwartz's data, and the impact of selecting the intervals over which to average by prior analysis of an interrelated data set. Julia Campos, Ericsson, and Hendry (1990) address these issues and show that (i) the aggregation from annual data to phase averages entails a loss of information by retaining only a subset of a previously averaged annual data set, (ii) phase-averaging does not materially reduce the serial correlation in Friedman and Schwartz's data, and (iii) phase-averaging involves a data-based set of filters but ignores the statistical consequences of such filters on later inferences. Figure 6 illustrates points (i) and (ii) for the time-series on velocity by jointly plotting v_t and \bar{v}_j . Neither the variance nor the autocorrelation of \bar{v}_j is notably smaller than that of v_t , whereas period-to-period changes in \bar{v}_j are much larger than those in v_t due to the observations being a phase rather than a year apart.

Thus, because of these unsatisfactory characteristics of phase-averaging, we analyze Friedman and Schwartz's original annual observations in an attempt to develop a money-demand equation that encompasses (1) and accounts for its failures. We next summarize the statistical framework and the associated theory of testing on which both the above analysis and the later models are based.

III. Econometric Modeling and Economic Theory

This section discusses the nature of empirical econometric models and serves as a guide to our uses of test statistics. Modeling is seen as an attempt to characterize data properties in simple parametric relation-

ships that are interpretable in light of economic knowledge, remain reasonably constant over time, and account for the findings of preexisting models. For general expositions of the associated methodology, see Hendry and Richard (1982, 1983), Hendry (1983, 1987), Hendry and Kenneth Wallis (1984), Ericsson and Hendry (1985), Christopher Gilbert (1986), and Aris Spanos (1986).

The formal methodology is based on the statistical theory of data reduction. Empirical models arise from *transformations* and *reductions* of the data generation process (DGP, a shorthand for the actual mechanism that generated the observed data), which is characterized by the joint density of the observable variables. The main steps producing models from the DGP are aggregation, algebraic transformation of data, marginalization, sequential conditioning, contemporaneous conditioning, truncation of lag length, and linearization. Each step entails a corresponding reduction and/or transformation of the parameters of the DGP to produce the reduced reparameterization in the econometric model. As a consequence, *all empirical models are derived entities*. Additionally, given the observed data and a formal model specification, the model's error process must contain everything in the data not explicitly allowed for by the model, and hence it also is derived rather than autonomous. This implies that models can be designed to satisfy preassigned criteria, as when residual autocorrelation is removed by a Cochrane-Orcutt transformation. While this example may be optimal in some circumstances, in others it entails invalid restrictions (cf. Hendry and Mizon, 1978).

The applied methodology distinguishes between the constructive and destructive roles of econometrics. The former corresponds to John Herschel's (1830) "context of discovery" and concerns issues of model design such as research efficiency. The latter is Herschel's "context of justification" and, for model evaluation, was illustrated in Section II. Criteria relevant to both design and evaluation are entailed by the conditions required to validate the (sometimes

implicit) sequence of reductions and transformations from which empirical models arise. Clearly, the validity of one-to-one data transformations is not an issue. However, reductions may entail a loss of information, and to judge that loss we use a taxonomy that partitions the universe of available information into disjoint sets: a) economic theory, the b) past, c) present, and d) future data of one's own model, e) the measurement system of the data, and f) the data of alternative models. These information sets generate the following criteria for design (in the context of discovery) and for evaluation (in the context of justification): A) theory consistency, B) innovation errors, C) weak

exogeneity, D) parameter constancy, E) data admissibility, and F) encompassing. A model is said to be *congruent* if it satisfies all of A–F and hence captures the salient features of the data and delivers reliable inferences on economic issues. Many of the criteria from A–F are standard and include the statistical and economic interpretation of estimated coefficients and the validity of a priori restrictions, goodness-of-fit and the absence of both residual autocorrelation and heteroscedasticity, valid exogeneity, predictive ability and parameter constancy, appropriate functional form, and the ability of a model to account for the properties of alternative models.

TABLE 2—CRITERIA FOR EVALUATING AND DESIGNING ECONOMETRIC MODELS

Null hypothesis	Alternative hypothesis	Statistic	Sources
B	first-order residual autocorrelation	dw	James Durbin and Geoffrey Watson (1950, 1951)
B	q th-order residual autocorrelation	$\xi_2[q];$ $\eta_2[q, T - k - q]$	George Box and David Pierce (1970); Leslie Godfrey (1978), Andrew Harvey (1981 p. 173)
B	q invalid parameter restrictions	$\eta_3[q, T - k - q]$	J. Johnston (1963 p. 126)
B	q th-order ARCH	$\xi_4[q]; \eta_4[q, T - k - 2q]$	Engle (1982)
B	skewness and excess kurtosis	$\xi_5[2]$	Jarque and Bera (1980)
B	heteroscedasticity quadratic in regressors (q quadratic terms)	$\eta_6[q, T - k - q - 1]$	White (1980), Desmond Nicholls and Adrian Pagan (1983)
B	q th-order RESET	$\eta_7[q, T - k - q]$	J. B. Ramsey (1969)
C	q instrumental variables not independent of errors	$\xi_8[q - k];$ $\eta_8[q - k, T - q]$	J. Denis Sargan (1958, 1964); Sargan (1980a p. 1136)
D	parameters not constant over subsamples	$\eta_9[k, T - 2k]$	R. A. Fisher (1922), Chow (1960 pp. 595–602)
D	predictive failure over a subset of q observations	$\eta_1[q, T - k - q]$	Chow (1960 pp. 594–5)

Notes: Null hypotheses correspond to criteria for design, as designated in the text. There are T observations and k regressors in the model under each null hypothesis. The value of q may differ across statistics, as may those of k and T across models and samples. $\xi_i[q]$ and $\eta_i[q, r]$ denote statistics that have central $\chi^2_{[q]}$ and $F_{[q, r]}$ distributions, respectively, under a common null and against the i th alternative. Thus, $\xi_2[q]$ and $\eta_2[q, T - k - q]$ both test for q th-order residual autocorrelation. We have labeled the Chow statistic $\eta_1[q, T - k - q]$ both to highlight the preeminence of the issue of constancy in the substantive debate on monetary behavior and because of its crucial role as an indirect test of weak exogeneity through testing the conjunction of hypotheses embodied in super exogeneity. The covariance test statistic $\eta_9[k, T - 2k]$ is often (and confusingly) referred to as the "Chow statistic" although Chow (1960 p. 592) was well aware of its presence in the literature.

In the context of discovery, satisfying a given criterion implies no loss of information from the associated reduction. In the context of evaluation, the criteria are interpretable as null hypotheses, and Table 2 summarizes the associated statistics. Since satisfying these criteria is a necessary condition for an empirical model to be congruent, failing any one of them is a sufficient condition for rejection. Conversely, since the route chosen for discovery cannot affect the *intrinsic validity* (or otherwise) of the finally selected model, issues of model design only concern the *efficiency* of alternative model-building strategies. Thus, we emphasize both explicitly *designing* empirical economic models to be congruent and rigorously *testing* the congruency of given model specifications. Since the latter takes models as formulated by their proprietors, the evaluative role of econometrics is not impugned by any of the current methodological debates (cf. Christopher Sims, 1980; Edward Leamer, 1983; Michael McAleer, Adrian Pagan, and Paul Volker, 1985; Hendry and Mizon, 1990). The failure to emphasize evaluation and testing is a major lacuna in Leamer's (1983) analysis, and our paper is a counterexample to the view that data evidence is not able to discriminate between alternative hypotheses. Rigorous evaluation of empirical claims seems a necessary first step toward taking the con out of economics, and indeed, Halbert White (1988) has shown that a sufficiently thorough testing procedure will arrive at a well-specified characterization of the DGP with confidence approaching certainty as the sample size grows without bound. The remainder of this section discusses the above criteria, their interrelationships, and their role in progressive research strategies.

A. Theory Consistency

Economic theory often suggests long-run relationships between economic variables, such as between two variables Y and Z (e.g., money and income) of the form $Y = KZ$ where K is a constant. In logs, that theory becomes $y = \kappa + z$ with $\kappa = \ln(K)$. *Cointegration* links the economic notion of a

long-run relationship between variables with a statistical model of those variables.⁹ For convenience, we adopt the time-series notation that a nonstationary variable integrated of order d [i.e., $I(d)$] requires differencing d times to make it stationary [i.e., $I(0)$]. For $I(1)$ variables y_t and z_t , arbitrary linear combinations $y_t + \delta z_t = u_t$ are also $I(1)$. If a value of δ exists such that u_t is $I(0)$, then y_t and z_t are "cointegrated" and so do not drift too far apart (e.g., because of agent behavior). Proportionality between Y and Z implies $\delta = -1$. Cointegration of y_t and z_t can be tested by testing for the absence of a unit root in u_t , that is, whether u_t is $I(0)$ rather than $I(1)$, with $I(0)$ being necessary for theory consistency.

B. Innovation Errors

Dynamic specification influences model design because of the statistical and economic importance of (white-noise) innovation errors. As a rule, dynamic misspecification invalidates inference, so dynamics cannot be safely ignored. There are many possible dynamic formulations for economic models (e.g., see Hendry, Pagan, and J. Denis Sargan, 1984); but since cointegration implies and is implied by the existence of a (dynamic) error-correction representation of the relevant variables, we consider the properties of and intuition behind error-correction models (cf. Sargan, 1964; James Davidson et al., 1978 pp. 679–82; Mark Salmon, 1982). For expositional simplicity, suppose that only the current values of y and z and their one-period lags matter, that y and z are cointegrated, and that the relationship is linear, in which case

$$(4) \quad y_t = \alpha + \alpha_1 y_{t-1} + \beta_0 z_t + \beta_1 z_{t-1} + \nu_t$$

$$\nu_t \sim \text{IN}(0, \sigma_\nu^2)$$

where IN means "is distributed indepen-

⁹See Clive Granger (1981), Hendry (1986), Engle and Granger (1987), Jeff Hallman (1987), Peter Phillips (1987), and James Stock (1987). Engle (1987) and Juan Dolado and Tim Jenkinson (1987) survey the literature.

dently normally." In (4), long-run homogeneity between y and z requires that $\alpha_1 + \beta_0 + \beta_1 = 1$. Rewriting the autoregressive-distributed lag relationship in (4) by imposing that restriction gives

$$(5) \quad \Delta y_t = \alpha + \beta \Delta z_t + \gamma(y_{t-1} - z_{t-1}) + \nu_t$$

where α , β , and γ are the corresponding *unrestricted* parameters.¹⁰ Intuitively, the term $\beta \Delta z_t$ reflects the immediate effect of a change in z_t on y_t . The term $\gamma(y_{t-1} - z_{t-1})$ (with γ negative for dynamic stability and, therefore, cointegration) is statistically equivalent to having $\gamma(y_{t-1} - \kappa - z_{t-1})$ instead, and hence reflects the effect on Δy_t of having y_{t-1} out of line with $\kappa + z_{t-1}$. Such discrepancies could arise from errors in agents' past decisions, with the presence of $\gamma(y_{t-1} - z_{t-1})$ reflecting their attempts to correct such errors; so, (5) belongs to the class of *error-correction* models (ECM's).¹¹ For a steady-state growth rate of Z_t equal to g (i.e., $g = \Delta z_t = \Delta y_t$) and $\nu_t = 0$, then solving (5), we have

$$(6) \quad Y_t = Z_t \exp\{[-\alpha + g(1 - \beta)]/\gamma\}$$

which reproduces the nonstochastic steady-state theory of proportionality between Y_t and Z_t . This example is readily extended to include further lags, nonproportionality, and nonlinearity, and so is representative of a large class of models that satisfy steady-state economic-theoretic restrictions and allow for general dynamic responses.

The choice of normalization of the cointegrating vector is an unresolved issue, but both economics and the data can help. Theory may suggest which variables agents aim to control and on which ones they condition their plans; and as parameter constancy in an empirical model is *not* invariant to nor-

malization when the economy exhibits structural change, data can preclude some choices. Thus, normalization and conditioning both lead to the next two issues: exogeneity and parameter constancy.

C. Weak Exogeneity

The four distinct concepts of exogeneity (weak, strong, super, and strict), discussed by Engle et al. (1983), correspond to different notions of being "determined outside the model under consideration" according to the purposes of the inferences being conducted (i.e., conditional inference, prediction, policy analysis, and forecasting, respectively).¹² In no case is it legitimate to "make variables exogenous" simply by not modeling them. Weak exogeneity can occur when agents act contingently on available information. If agents use that information efficiently, innovation errors are implied, relating back to the issue of dynamic specification. Weak exogeneity is testable, often as an implication of super exogeneity (and so of models having constant parameters). Section IV further discusses exogeneity in the context of the empirical model.

D. Parameter Constancy

Parameter constancy is at the heart of model design from both statistical and economic perspectives. Since economic systems are far from being constant and the coefficients of derived ("nonstructural" or "reduced form") equations may change when any of the underlying parameters or data correlations change, it is important to identify empirical models that have reasonably constant parameters, which remain interpretable when change occurs. As seen in Section II above and Section IV below, recursive estimation provides an incisive tool for investigating parameter constancy, both through the sequence of estimated coeffi-

¹⁰With the lag operator L defined as $Lx_t = x_{t-1}$, we let the difference operator Δ be $(1 - L)$; hence, $\Delta x_t = x_t - x_{t-1}$. More generally, $\Delta^q x_t = (1 - L)^q x_t$. If q (or r) is undefined, it is taken to be unity.

¹¹Alternatively, Stephen Nickell (1985) justifies error-correction mechanisms as arising from the optimal response of economic agents in some dynamic environments.

¹²This formulation is also discussed in Jean-Pierre Florens and Michel Mouchart (1985a, b) and builds on Tjalling Koopmans (1950) and Ole Barndorff-Nielsen (1978).

391
001

P 7831

cient values and via the associated Chow statistics for constancy.¹³ The Chow statistics also play crucial roles for testing weak exogeneity indirectly through testing the conjunction of hypotheses embodied in super exogeneity and for testing feedback versus feedforward empirical models (cf. Engle et al., 1983; Hendry, 1988).

E. Data Admissibility

Many economic variables are inherently positive, a model property ensured by the use of logarithmic transformations of the data. However, if the resulting model does not correspond to the DGP, the cost of enforcing data admissibility may be a loss of parameter constancy.

F. Encompassing

This concept can be understood intuitively as follows. Suppose model 1 predicts $\hat{\theta}$ as the value for the parameter θ in model 2, while model 2 actually has the estimate $\hat{\theta}$. Model 1 encompasses model 2 if $\hat{\theta}$ is "statistically close" to $\hat{\theta}$, so that model 1 explains why model 2 obtains the results it does. To the extent that model 1 accurately mimics the DGP, it will encompass model 2. For single equations estimated by least squares, a necessary condition for encompassing is *variance dominance*, where one equation variance-dominates another if the former has a smaller error variance.¹⁴ Thus, encompassing is more demanding than selecting models purely on the basis of their goodness of fit. It is also consistent with the concept of a progressive research strategy (e.g., see Imre Lakatos, 1970; Alan

Chalmers, 1976), since an encompassing model is a "sufficient representative" of previous empirical findings.¹⁵

These six criteria not only characterize the conditions required to sustain reductions, they also correspond to concepts central to econometric analysis. Taking the reductions in turn, a set of current-dated variables can be eliminated (marginalized) without loss of information if those retained correspond to *sufficient statistics*, and lagged variables can be marginalized if they do not *Granger-cause* the remaining variables. Sequential conditioning generates *innovation errors*, and contemporaneous conditioning is valid if the variables so treated are weakly *exogenous* for the parameters of interest. The concept of *encompassing* introduced above determines the limits to model reduction (i.e., the degree of *parsimony* feasible). As shown in Hendry and Richard (1989), parsimonious encompassing (in which the smaller model accounts for the results of a larger model within which it is nested) is transitive, antisymmetric, and reflexive, thus defining a partial ordering over models and thereby sustaining a progressive research strategy.

IV. Econometric Modeling of Money Demand Using the Annual Data

This section develops a conditional econometric model of money demand on the annual data series for 1878–1970 only ($T = 93$), since the data from 1971 to 1975 appear to have a different stochastic structure [see (11) below]. We follow the procedures outlined in Section III, representing the joint density of $(m_t, p_t, i_t, RS_t, RL_t)$ in terms of an autoregressive-distributed lag model for m_t conditional on (p_t, i_t, RS_t, RL_t) and a marginal model for (p_t, i_t, RS_t, RL_t) . The conditional model is simplified to an ECM and evaluated in light of

¹³These tests of constancy are intimately related to tests of forecast accuracy (cf. R. Brown, Durbin, and J. M. Evans, 1975; Hendry, 1979; Jan Kiviet, 1987). Jean-Marie Dufour (1982) elegantly summarizes recursive techniques and their implications.

¹⁴Formally, variance dominance refers to the underlying (and unknown) error variances. Without loss of clarity, we often will say a model variance-dominates another if the *estimated residual* variance of the former is smaller than that of the latter.

¹⁵For comprehensive accounts of tests for encompassing and of related nonnested hypothesis tests, see Mizon and Richard (1986), Mizon (1984), James MacKinnon (1983), and Hashem Pesaran (1982).

the model design criteria.¹⁶ Phillips (1988) and Phillips and Bruce Hansen (1990) demonstrate that this error-correction approach produces nearly optimal inferences in cointegrated processes. Such a methodology of "learning from the data" while being guided by economic theory in the interpretation of results contrasts with the approach adopted by Friedman and Schwartz of using regression results to corroborate their economic theory. We conclude this section by considering two important issues for economic policy: the constancy of the money-demand function and the exogeneity of money.

The economic framework of our empirical model is that of a log-linear long-run money-demand function:

$$(7) \quad (m - p)^* = \delta_0 + \delta_1 i - \delta_2 RS - \delta_3(1 + \dot{p})$$

where an asterisk denotes the long-run target value and \dot{p} is the rate of inflation (cf. Friedman, 1956). Equation (7) parallels the condition $y = \kappa + z$ in Section III-A. Dynamic adjustment is characterized by a contingent planning model of the form

$$(8) \quad \Delta(m - p)_t = \lambda_0(L)\Delta(m - p)_{t-1} + \lambda_1(L)\Delta p_t + \lambda_2(L)\Delta i_t + \lambda_3(L)\Delta rs_t + \lambda_4(L)\Delta rl_t + \lambda_5[(m - p)_{t-1}^* - (m - p)_{t-1}] + \varepsilon_t$$

where $\lambda_i(L)$ ($i = 0, \dots, 4$) denotes a finite polynomial in the lag operator L and ε_t is the deviation of the outcome from the plan. This approach generalizes the conventional partial-adjustment model, allowing separate reaction speeds to the different determi-

nants of money demand (reflecting potentially different costs of adjustment and of disequilibrium), yet via the error-correction mechanism ensures that long-run targets are achieved (e.g., velocity and interest rates are cointegrated) (cf. Hendry et al., 1984). Economically, (8) is related to the theory of money adjustment in Merton Miller and Daniel Orr (1966) and Ross Milbourne (1983), in which the short-run factors determine money movements *given* the desired bands, and the longer-run factors influence the levels of the bands (e.g., see Gregor Smith, 1986). To be interpretable as a demand equation, $\lambda_1(1) \leq 0$, $\lambda_2(1) \geq 0$, $\lambda_3(1) \leq 0$, and $\lambda_4(1) \leq 0$; and for cointegration, $\lambda_5 < 0$. However, the choice of parameterization (e.g., in terms of lagged first differences or higher-order differences) is arbitrary *within* lag polynomials, so no sign restrictions on individual polynomial coefficients can be imposed a priori. Moreover, since monetary theory does not yet specify how quickly agents react to changes in real time, the orders of the $\lambda_i(L)$ must be data-based.

Statistically, (8) is an ECM reparameterization of an autoregressive-distributed lag model of the variables in levels, as is (5) of (4). A model such as (8) is of interest only if the regressors are weakly exogenous for the resulting parameters, which in turn must be constant and invariant to historical changes in the processes determining the marginal distributions. The error in (8) will be an innovation if the plan efficiently incorporates available information and the model is correctly specified. All of these issues are examined below in order to discriminate between contingent planning and expectations interpretations of the empirical model.

In Hendry and Ericsson (1983), we presented an autoregressive-distributed lag representation of money conditional on prices, incomes, and interest rates to establish the innovation error variance. That formulation was simplified to an ECM like (8) based on extant money-demand models for the United Kingdom, with a static-equilibrium solution of the form $v = -\delta_0 + \delta_2 RS$ (thus taking v_t and RS_t to be cointegrated) and an equation standard error of

¹⁶See Hendry and Mizon (1978) and McAleer et al. (1985) on modeling from general to simple; see Leamer (1978) for an analysis of specification searches.

TABLE 3—A GENERAL AUTOREGRESSIVE-DISTRIBUTED LAG REPRESENTATION FOR MONEY (m_t),
CONDITIONAL ON INCOMES, PRICES, AND INTEREST RATES

Variable	lag i (or index)						$\Sigma_{i=0}^5$
	0	1	2	3	4	5	
m_{t-i}	-1.0 (0.0)	1.316 (0.157)	-0.621 (0.226)	0.295 (0.242)	-0.091 (0.220)	0.00047 (0.111)	-0.100 (0.084)
p_{t-i}	0.447 (0.077)	-0.410 (0.120)	0.201 (0.119)	-0.176 (0.116)	0.083 (0.110)	-0.050 (0.074)	0.095 (0.093)
i_{t-i}	0.087 (0.072)	0.031 (0.121)	0.047 (0.090)	-0.024 (0.088)	0.034 (0.091)	-0.065 (0.071)	0.110 (0.078)
rs_{t-i}	-0.019 (0.00908)	0.014 (0.016)	-0.00417 (0.010)	0.00510 (0.011)	-0.00449 (0.011)	0.00407 (0.00889)	-0.00403 (0.013)
rl_{t-i}	-0.069 (0.045)	-0.075 (0.070)	0.147 (0.072)	-0.084 (0.075)	0.071 (0.073)	-0.00664 (0.052)	-0.016 (0.041)
\hat{u}_{t-1}^i		0.077 (0.092)	0.485 (0.204)	-1.821 (1.285)			
D_i		3.993 (1.220)	0.607 (1.751)	3.624 (1.024)			
Constant	-0.201 (0.137)						

Notes: Values in parentheses are estimated standard errors; $T = 1878-1970$, $R^2 = 0.99987$, $\hat{\sigma} = 1.5535$, $dw = 1.94$, $\xi_2[11] = 9.62$, $\eta_2[3, 54] = 0.46$, $\eta_4[4, 49] = 1.39$, $\xi_5[2] = 1.01$, $\eta_6[39, 17] = 0.23$, $\eta_7[2, 55] = 1.28$. For readability, coefficients and estimated standard errors on D_1 , D_2 , and D_3 are multiplied by 100.

1.71 percent of M .¹⁷ Those results stimulated further studies, including an improved specification on the same information set in Andrew Longbottom and Sean Holly (1985), a nonlinear reformulation in Alvaro Escribano (1985), and an extended information set for 1875-1913 in Jan Klovland (1987).¹⁸ Longbottom and Holly's and Escribano's models were significant improvements on our 1983 model (i.e., each of their models encompassed ours, but ours could not encompass either of theirs). However,

¹⁷Hendry and Ericsson (1985) investigate the hypothesis that v_t is a random walk and conclude that there is little contrary evidence. Likewise, the data on m , i , p , $m-p$, rs , rl , and p^* all behave like $I(1)$ series, whereas their corresponding first differences behave like $I(0)$ series. Nevertheless, equations (9) and (10) show that a more general cointegrating relationship can be established.

¹⁸Klovland (1987) constructs a new measure of the own interest rate on M .

neither of their models could encompass certain features of the other, indicating that an improved specification might be possible.

Continuing in a progressive research strategy, we utilize their (and our previous) evidence, beginning with the static cointegration regression for v_t and RS_t :

$$(9) \quad (\overline{m-i-p})_t = -0.309 - 7.00RS_t$$

($T = 1873-1970$, $R^2 = 0.56$, $\hat{\sigma} = 10.86$ percent, $dw = 0.33$, $ADF(1) = -2.77$). (Adding RL makes little difference, unsurprisingly if RS and RL are cointegrated.) Exact significance levels are not available for David Dickey and Wayne Fuller's (1979, 1981) augmented statistic $ADF(1)$ or for Sargan and Alok Bhargava's (1983) Durbin-Watson-based statistic when testing for a unit root in the residuals, but the 10-percent critical values in Engle and Granger (1987), based on simulation, are respectively -2.84

and 0.32 for 100 observations, leading to an inconclusive outcome. Engle and Granger (1987) show that these tests have low power against highly dynamic stationary alternatives, so we cautiously proceed under the assumption of cointegration. Economically, the solution to (9) is within the framework of (7) and implies that a one-percentage-point increase in the short-term interest rate (e.g., from 5 percent to 6 percent) reduces M relative to PI by 7 percent in the long run.

Cointegration implies an error-correction representation, so to model the behavior of money we estimate an unrestricted fifth-order autoregressive-distributed lag equation related to (4); it is shown in Table 3 together with relevant test statistics.¹⁹ The model in Table 3 generalizes on (4) by inter alia allowing the speed of adjustment to vary with the extent of disequilibrium via nonlinear error-correction terms. Following Escribano (1985), Table 3 includes the lagged level, square, and cube of \hat{u}_t , the residual from (9). Three zero-one dummies (D_1 , D_2 , and D_3), which are unity for 1914–1918, 1921–1955, and 1939–1945, respectively, also appear. This regression satisfies all of the diagnostic checks reported and provides generally sensible estimates, despite very high intercorrelations between regressors. Most coefficients of lags beyond three are negligible as well as insignificant, and deleting those lags lowers $\hat{\sigma}$. The entailment static solution is consistent with long-run price and income elasticities of around unity.

The representation in Table 3 was simplified to the error-correction model (10) using the approach described in Hendry (1983) of first transforming the model to an interpretable and near orthogonal specification paralleling (8) and then eliminating negligi-

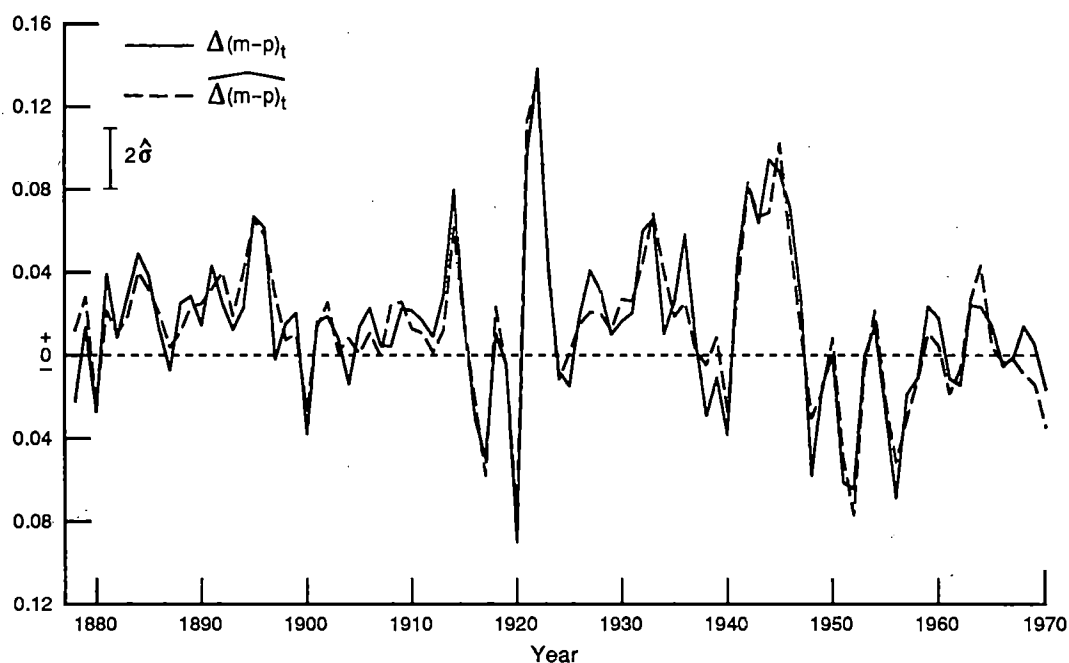
ble and insignificant effects:

$$\begin{aligned}
 (10) \quad & \overline{\Delta(m-p)}_t \\
 & = 0.45 \Delta(m-p)_{t-1} \\
 & \quad [0.06] \\
 & \quad - 0.10 \Delta^2(m-p)_{t-2} \\
 & \quad [0.04] \\
 & \quad - 0.60 \Delta p_t + 0.39 \Delta p_{t-1} \\
 & \quad [0.04] \quad [0.05] \\
 & \quad - 0.021 \Delta r_t - 0.062 \Delta^2 r_t \\
 & \quad [0.006] \quad [0.021] \\
 & \quad - 2.55(\hat{u}_{t-1} - 0.2)\hat{u}_{t-1}^2 \\
 & \quad [0.59] \\
 & \quad + 0.005 + 3.7(D_1 + D_3)_t \\
 & \quad [0.002] \quad [0.6]
 \end{aligned}$$

($T = 1878-1970$, $R^2 = 0.87$, $\hat{\sigma} = 1.424$ percent, $\eta_2[6, 78] = 1.56$, $\eta_3[27, 57] = 0.50$, $\eta_4[4, 76] = 1.24$, $\xi_5[2] = 1.6$, $\eta_6[15, 68] = 0.87$, $\eta_7[2, 82] = 0.21$; see Table 2 for definitions of the statistics). Values in brackets are heteroscedasticity-consistent estimated standard errors (see White, 1980; Desmond Nicholls and Pagan, 1983; MacKinnon and White, 1985).

Concerning its economic interpretation, (10) is similar in form and in numerical parameter values to several successful money-demand models for the United Kingdom (cf. Hendry and Mizon, 1978; Hendry, 1979; John Trundle, 1982; Davidson, 1987; Keith Cuthbertson, 1988). Its coefficients satisfy the sign restrictions on the $\lambda_t(1)$ in (8) to be interpretable as a money-demand function. The coefficients' sizes imply large immediate responses to changes in inflation and interest rates but slow adjustment subsequently to remaining disequilibria, via the error-correction term. Inflation enters as $\Delta p_t + \Delta^2 p_t$ (approximately), which is a predictor of next period's inflation, optimal if prices vary quadratically. Thus, (10) has a forward-looking interpretation, albeit one based on *data functions* rather than on

¹⁹We have chosen five lags on the grounds that this implies 36 parameters estimated in Table 3 (vs. 93 observations total), in line with guidelines in Sargan (1980b p. 880). Also, five lags implies a maximum lag of approximately one cycle.

FIGURE 7. EQUATION (10): ACTUAL AND FITTED VALUES FOR $\Delta(m-p)_t$

models of the right-hand-side variables.²⁰ The coefficients correspond to nearly orthogonal decision variables (with only two of the 28 regressor intercorrelations exceeding 0.5), consistent with (10) representing a contingent plan of agents who partition available information into conceptually separate entities.

The empirical parameterization in (10) exhibits multiple equilibria, with two corresponding to the long-run solution (9) and a third being that solution shifted by 20 percent. In that sense, the results are consistent with the use of an adjustment factor of about that order by Friedman and Schwartz. However, the dating of the disequilibria is determined by the values of \hat{u}_{t-1} and operates over the entire sample, not just for the period 1921–1955.

Concerning the statistical attributes of (10), the various diagnostic checks are in-

significant (if regarded as test statistics) and indicate design of a model congruent with the information available. Even so, the range of alternatives considered is sufficiently large to endow (10) with some credibility. From η_2 , the residuals are white noise and, from η_3 , are also an innovation process against the information set in Table 3. There is no ARCH, RESET, or heteroscedastic evidence of misspecification; the residuals are approximately normally distributed; and (10) encompasses our earlier models and those of Escobano and of Longbottom and Holly (but not conversely).²¹ Figure 7 shows the

²⁰Campos and Ericsson (1988) propose this interpretation for similar inflation terms entering an equation for consumers' expenditure in Venezuela.

²¹By appropriately (statistically) reducing the density for our model, one could derive an estimate of the parameters in Friedman and Schwartz's model. From that constructed estimate and the estimate they actually obtain, one could test whether our model (using annual data) encompasses theirs (using phase-average data). Note also that it is easy to construct examples for which the parameters in their model would not be constant over time but those in ours would be [e.g., the parameters in (10) are constant, p_t is Granger-caused by m_t , and the coefficients in that price equation change over time]. Existence of the converse would

actual and fitted values for the rate of change in real money over the sample period. Visual comparison of Figures 4 and 7 (noting the logarithmic scale in both) highlights the better fit of the annual model: the error variance of (10) is less than one-tenth of that in (1). Although (10) has a rate of change as the dependent variable, it is an equation in log-levels because of the error-correction term, as shown in (4)–(5) above, so direct comparison of the two graphs is valid.

The two remaining issues are constancy and exogeneity. Any claim to the constancy of a model for money demand would need both constant parameters and a similar goodness of fit over each of the epochs described above for Figure 1; in Section II, we demonstrated that (1) has neither.²² To investigate the constancy of our model (10), we adopt the recursive estimator, since the one-step innovations allow the construction of sequences of constancy tests. Given the world wars dummy ($D_1 + D_3$), 1915 is the earliest date for a continuous sequence until 1970, although a separate exercise is possible for the subsample 1878–1913 (see Fig. 11, below). Graphical presentation is efficient for reporting the large volume of evaluation output: Figure 8 records the one-step residuals and the corresponding calculated equation standard errors (i.e., $y_t - \hat{\beta}'_t x_t$ and $0.0 \pm 2\hat{\sigma}_t$ in a standard notation). *It is visually apparent that $\hat{\sigma}$ has varied little over the 56-year test period.* Further, none of the Chow statistics for the sequence 1915, 1915–1916, 1915–1917, ..., 1915–1970 or the sequence 1915–1970, 1916–1970, 1917–1970, ..., 1969–1970, 1970 is significant at even the 5-percent level. Figures 9

and 10 show the numerical values of two central coefficients, namely, those for Δp_t and the error-correction $(\hat{u}_{t-1} - 0.2)\hat{u}_{t-1}^2$, together with plus-or-minus twice their sequentially estimated standard errors, which provide an approximate 95-percent confidence interval.²³ Other than a minor fluctuation directly following World War I, the former varies by only a tiny fraction of its *ex ante* standard error; the latter is highly significant for the entire sample and also varies little relative to the estimated uncertainty. In both cases, the accrual of information is apparent from the reduction in the width of the confidence interval over time. Figure 11 graphs the one-step residuals for 1892–1913, using only 15 observations to initialize estimation; not only is $\hat{\sigma}$ constant, but its value throughout is very close to that for the full sample. In brief, our *conditional* reformulation both fits well and is constant over the century to 1970, even though the time-aggregated phase-average data reveal the nonconstancy of (1) despite the attendant loss of information.

Thus, we reach the issue of the exogeneity of money. Our analysis has taken every contemporaneous variable other than m as if it were weakly exogenous (i.e., that it is valid to condition upon them for purposes of statistical inference), interpreting the coefficients of the resulting model as those of a money-demand equation; by construction, (10) is *invariant* to whether Δm_t or $\Delta(m - p)_t$ is the regressand. The constancy of (10) reinforces our interpretation, which is also consistent with the institutional structure of U.K. money markets in which the money stock appears to be endogenously determined by the decisions of the private sector since the Bank of England in effect acts as a lender of the first resort by standing ready to rediscount first-class bills at the going Bank Rate or Minimum Lending Rate (see e.g., R. Hawtrey, 1938; Congdon, 1983; Goodhart, 1984). Its constancy suggests that the conditioning variables may be super ex-

immediately refute any claim to encompass their findings.

²²A possible objection might be that a money-demand equation's "constancy" need not be precise but only relatively better than (say) the consumption function's constancy. Since Friedman and Schwartz assert that their model is indeed constant, such an objection is not germane. However, in response to Mayer (1982 p. 1534), we note that the U.K. consumption function was investigated in Hendry (1983) and was shown to be remarkably constant both before and after World War II.

²³Note, however, that the coefficients within \hat{u}_{t-1} are full-period estimates, as justified by the distributional results in Engle and Granger (1987).

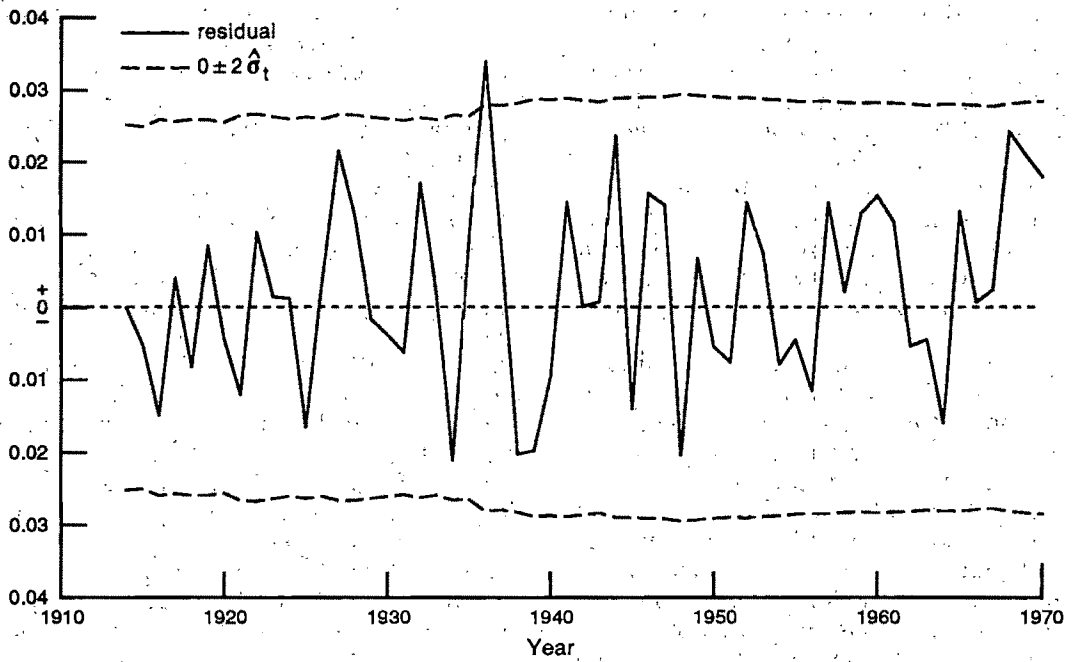


FIGURE 8. EQUATION (10): ONE-STEP RESIDUALS AND THE CORRESPONDING CALCULATED EQUATION STANDARD ERRORS

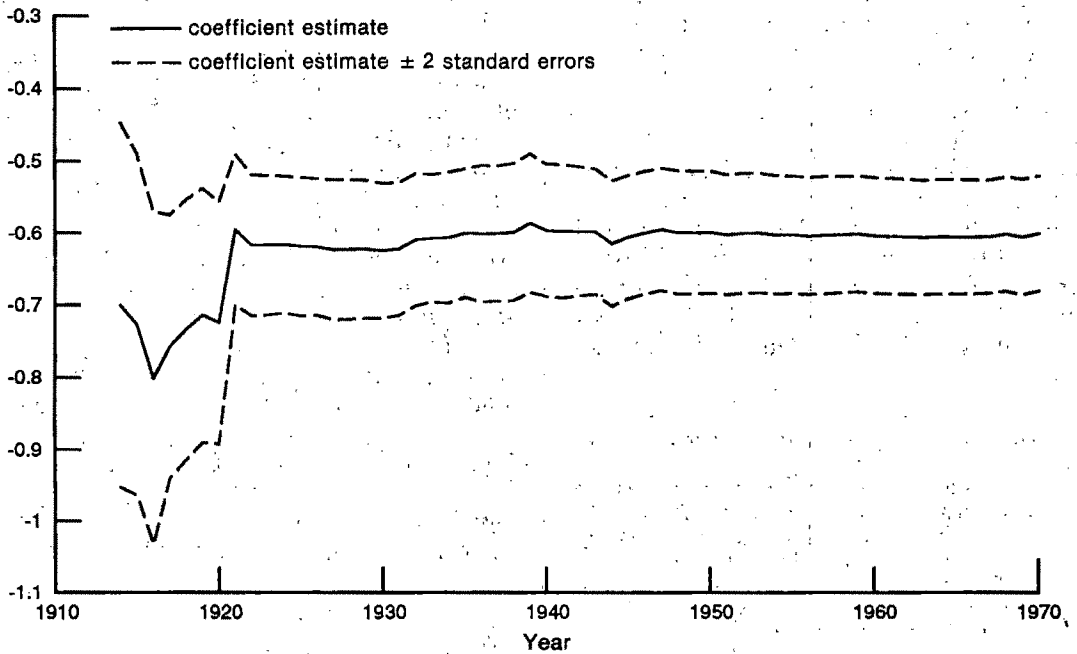


FIGURE 9. EQUATION (10): RECURSIVE ESTIMATION OF THE COEFFICIENT ON Δp_t AND ITS ESTIMATED STANDARD ERROR

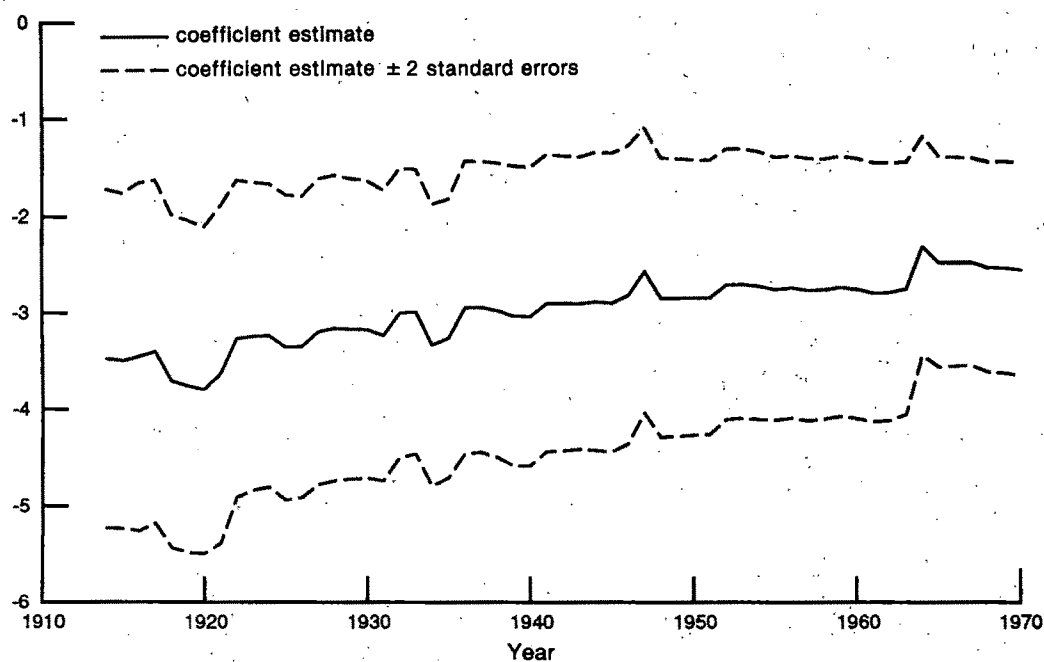


FIGURE 10. EQUATION (10): RECURSIVE ESTIMATION OF THE ERROR-CORRECTION COEFFICIENT AND ITS ESTIMATED STANDARD ERROR

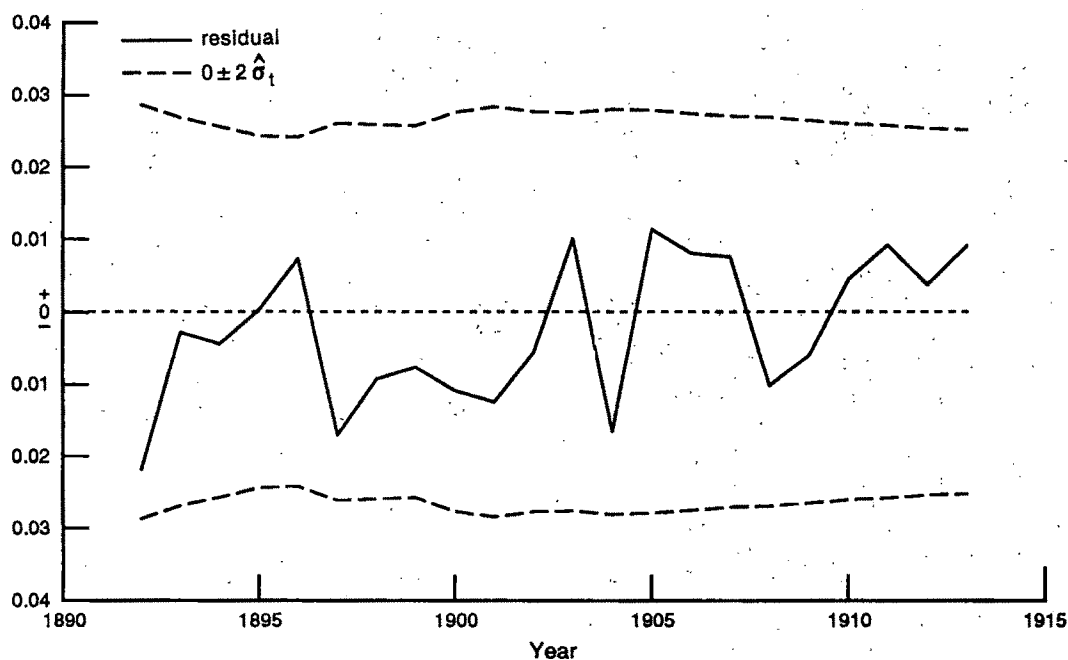


FIGURE 11. EQUATION (10) WITHOUT $(D_1 + D_3)_t$: ONE-STEP RESIDUALS AND THE CORRESPONDING CALCULATED EQUATION STANDARD ERRORS

ogenous for the demand parameters over the sample (i.e., that the parameters of the conditional model remain constant even though the DGP of the conditioning variables changes over the sample). The exogeneity of prices and endogeneity of money are substantive issues, noting that it is commonplace in macroeconomics to determine prices via the money-demand equation, taking income, interest rates, and an observed money supply as given and equating supply to demand (cf. Robert Barro, 1987 pp. 128–31, 195–210). In particular, the empirical super exogeneity of prices (shown below) invalidates “inverting” the money-demand equation to obtain prices. Engle and Hendry (1989) show that several implications of super exogeneity are testable.

First, direct tests of super exogeneity can be constructed, but they require additional observables, the relevance of which to (10) is that those observables should not have been used in model design [e.g., none of the variables dropped in simplifying from Table 3 to (10) would give the tests power]. For these data, the crucial issue is the weak exogeneity of Δp_t , since, if that is accepted, (10) cannot sustain the interpretation of determining prices with money exogenous. The additional instruments we have tried are a trend and current and lagged U.S. inflation (Δp_t^* and Δp_{t-1}^*), using a recursive instrumental variables estimator to investigate constancy and exogeneity conjointly (see Hendry and Adrian Neale, 1987). On endogenizing Δp_t , $\hat{\sigma}$ remains virtually unchanged at 1.44 percent, and Sargan’s (1958) statistic for testing the validity of the instruments yields $\xi_8[2] = 3.62$, justifying the choice of instruments and being consistent with the weak exogeneity of Δp_t .²⁴ More-

over, the demand equation is constant despite the manifest nonconstancy of the instrumenting equation (cf. Fig. 12), revealing how informative this data set is and precluding a forward-looking expectations interpretation of (10) (cf. Hendry, 1988).²⁵

Second, super exogeneity is not invariant to alternative factorizations of the joint density of money and prices when parameters in that density are not constant over time. That is, if the model (10) were inverted to make Δp_t the regressand, conditional on Δm_t , as an exogenous variable, then the resulting equation should be nonconstant. This is the case, as seen in Figure 13, which records the behavior of the estimated coefficient of Δm_t in the inverted equation. Further, the equation standard error becomes 2.6 percent and, unlike (10), the inverted equation cannot encompass the simple model that Δp_t^* determines Δp_t through a purchasing power parity condition: $\eta_3[1, 83] = 14.9$.

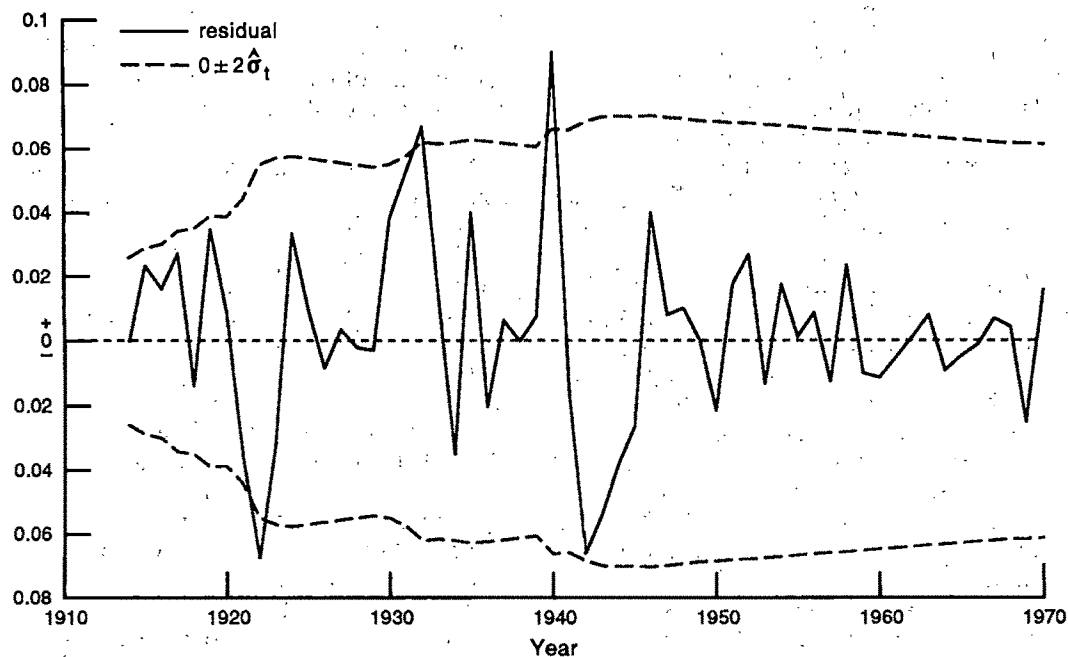
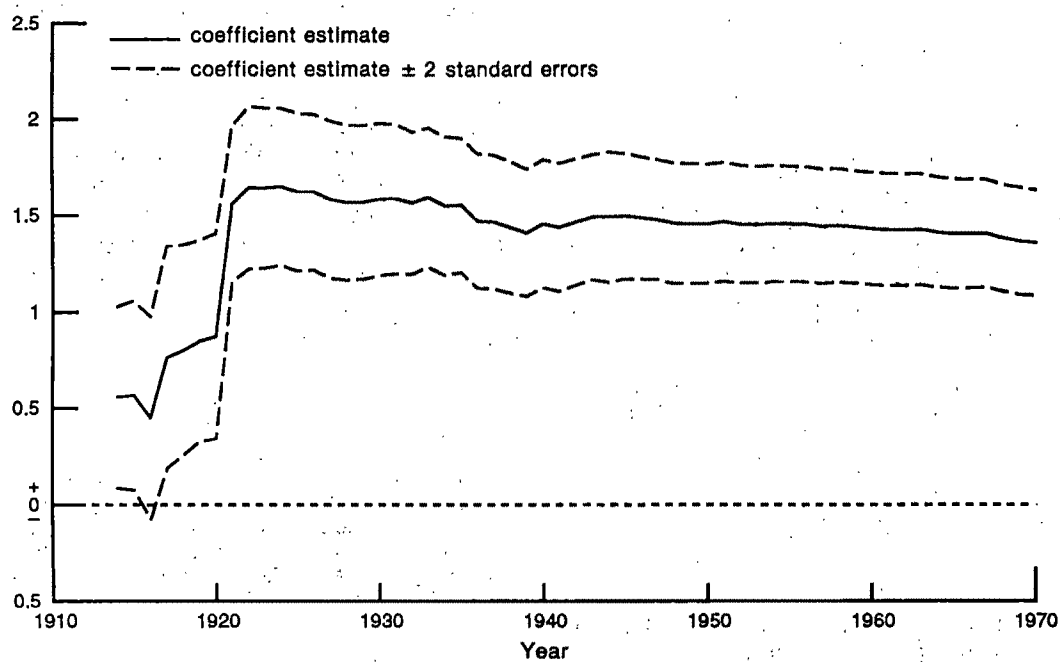
This evidence is inconsistent with the hypothesis that, over the period 1878–1970, exogenous money determined prices in the United Kingdom via a stable money-demand function, *precisely because we have established a constant money-demand model conditional on prices*. In Hendry and Ericsson (1986), we propose an alternative mechanism for money causing inflation: namely, the long-run effects of deviations from purchasing-power parity.

The final test of super exogeneity is that the parameters are invariant to regime changes: here, we exploit the joint introduction of floating exchange rates and Compe-

²⁴We also investigated the assumed *joint* weak exogeneity of prices and short- and long-term interest rates in (10). (The weak exogeneity of income is not at issue because it appears only at a lag, via the error-correction term.) When estimated with Δp_t^* , Δp_{t-1}^* , rs_{t-1} , rs_{t-2} , rl_{t-1} , rl_{t-2} , and a trend as instruments and treating $\Delta(m-p)_t$, Δp_t , Δrs_t , and $\Delta_2 rl_t$ as endogenous, $\hat{\sigma}$ increases only slightly to 1.56 percent, the coefficients remain virtually unchanged from (10), and Sargan’s (1958) statistic is $\xi_8[4] = 4.96$ (insignificant). Further, estimation of the reduced-form equations for

Δp_t , Δrs_t , and $\Delta_2 rl_t$ by recursive multivariate least squares reveals nonconstancy in each. If there were simultaneity bias in (10), the parameter estimates in (10) would change as those in the reduced-form equations did; thus, the *constancy* of (10) in spite of nonconstant reduced-form equations implies the weak exogeneity of prices and interest rates. See Hendry, Neale, and Srba (1988) on recursive multivariate least squares, and see Jean Bronfenbrenner (1953) on the relationship between simultaneity bias and reduced-form parameters.

²⁵See Engle (1984) for a survey of exogeneity tests and Kiviet (1985, 1987) for finite-sample evidence on their performance.

FIGURE 12. ONE-STEP RESIDUALS FROM THE INSTRUMENTING EQUATION FOR Δp_t USING Δp_t^* FIGURE 13. RECURSIVE ESTIMATION OF THE COEFFICIENT ON Δm_t IN THE INVERTED MODEL FOR Δp_t USING (10)

tion and Credit Control regulations in 1971. Narrow money demand (U.K. M_1 measure) apparently was not perturbed by these switches, but most investigators have found major parameter changes in models of broad money measures (e.g., U.K. $\text{£}M_3$; see Hendry and Mizon [1978] and Michel Lubrano, Richard Pierse, and Richard [1986] for overviews). A Chow test of (10) using predictions over 1971–1975 yields $\eta_1[5, 84] = 19.5$, and $\hat{\sigma}$ rises to 2.03 percent, rejecting constancy and super exogeneity. The growth rate of nominal money over this test period substantially exceeds that of any previous episode (including both world wars) and could reflect a disequilibrium adjustment to the removal of the chronic cartelization of the U.K. banking system (the motivation for the change in banking regulations). Indeed, Lubrano et al. (1986) find that the long-run relationship of money demand to prices and incomes holds after 1970 but that the short-run relationship is severely perturbed, with money demand *increasing* as competitive interest rates increase. With that structure in mind, we obtained

$$\begin{aligned}
 (11) \quad & \overline{\Delta(m-p)_t} \\
 & = 0.47\Delta(m-p)_{t-1} \\
 & \quad [0.06] \\
 & \quad - 0.11\Delta^2(m-p)_{t-2} - 0.59\Delta p_t \\
 & \quad [0.04] \quad [0.04] \\
 & \quad + 0.41\Delta p_{t-1} - 0.017\Delta rs_t \\
 & \quad [0.04] \quad [0.006] \\
 & \quad - 0.078\Delta_2 r1_t \\
 & \quad [0.019] \\
 & \quad - 1.15(\hat{u}_{t-1} - 0.2)\hat{u}_{t-1}^2 + 0.007 \\
 & \quad [0.19] \quad [0.002] \\
 & \quad + 3.4(D_1 + D_3)_t + 0.071(D_4)_t \\
 & \quad [0.6] \quad [0.010] \\
 & \quad + 0.090(D_4)_t \cdot \Delta rs_t \\
 & \quad [0.020]
 \end{aligned}$$

($T = 1878-1975$, $R^2 = 0.88$, $\hat{\sigma} = 1.478$ percent, $\eta_2[6, 81] = 1.19$, $\eta_4[2, 83] = 1.31$, $\xi_5[2] = 2.9$, $\eta_6[18, 68] = 0.43$), where D_4 is a dummy which is unity over the period 1971–1975

and zero otherwise, and where \hat{u}_{t-1} is still calculated from (9). Now $\hat{\sigma}$ and most of the coefficients other than those involving D_4 are virtually unaltered, consistent with Lubrano et al.'s results.

A potential explanation for this finding, consistent with the earlier evidence and economic analysis, is presaged by Klovland's (1987) result for pre-1914 data that the *own* interest rate on broad money is an important omitted variable from the present information set.²⁶ Over much of the sample, U.K. commercial banks acted like a cartel with administered (and generally low) deposit interest rates; this situation changed after 1970 due to the competition regulations. Thus, own interest rates rose rapidly, altering the historical differentials and inducing predictive failure in models that excluded that variable. We plan to extend Klovland's data set to test this conjecture and to continue the progressive research strategy by encompassing previous models.

V. Conclusions

This paper focuses on the evaluation of Friedman and Schwartz's (1982) empirical model for U.K. money demand and the design of an improved specification using their annual data. At the heart of model evaluation are the issues of model credibility and validity and the role of corroborating evidence. The failure by Friedman and Schwartz to present statistical evidence pertinent to their main claims about the United Kingdom leaves those claims lacking in credibility. The presence of substantial misspecification invalidates many of their infer-

²⁶In Hendry and Ericsson (1983 pp. 77–8), we experimented with RS replaced by RN, Friedman and Schwartz's (1982 p. 270) measure of the marginal cost of money [$RN \equiv RS \cdot (H/M)$]. The resulting estimates are very similar to those found using RS and reveal no improvement in the constancy of the interest-rate coefficient over 1971–1975. Following suggestions by Chris Pissarides and by Ross Starr (1982), we also experimented with adding variables that measured interest-rate volatility (e.g., using $0.2\sum_{i=0}^4 [RS_{t-i} - RS_t^*]^2$ where $RS_t^* = 0.2\sum_{i=0}^4 RS_{t-i}$). The effect was largest for the most recent period but did not produce constant parameters or a constant fit over 1971–1975.

ences from equations based on the phase-average data (cf. Table 1). In particular, their final money-demand equation is not constant, contrary to their claim; and, on testing assumptions such as price homogeneity and the absence of trends, rejection results. Such negative findings are consistent with those reported by Meghnad Desai (1981 [especially Ch. 4]). The procedure of averaging data over business-cycle phases did not notably reduce the serial correlation in the data series but did lose information, leading to rather badly fitting equations. As an alternative, we recommend analyzing the annual data and modeling "trend" and "cycle" jointly.

Corroborating a subset of the implications of a theory is not by itself an adequate justification for deeming the theory useful (see e.g., Friedman, 1953 pp. 8–9; Karl Popper, 1959 section 82; Lawrence Boland, 1982 Ch. 1). That is illustrated by the contrast between Friedman and Schwartz's claims to have empirically corroborated various aspects of their theories and our evidence that those claims are actually refutable from the same data. Only well-tested theories that have successfully weathered tests outside the control of their proponents and that encompass the gestalt of existing empirical evidence seem likely to provide a useful basis for applied economic analysis and policy.

The tests used to evaluate Friedman and Schwartz's models indicate that scope exists for model development but do not indicate what changes are required. We attempted an improved specification and presented an econometric model of the demand for money in the United Kingdom during 1878–1970 which incorporates economic theory, satisfies a wide range of statistical criteria, and encompasses existing models based on the same data set. Those features are essential for a model to characterize the underlying data generation process adequately. On a substantive level, the empirical constancy of our model is consistent with a structure in which nominal money is endogenously determined by demand factors, conditional on prices, incomes, and interest rates. Undoubtedly, the model pro-

posed above is not the end of the story since, for example, parameter nonconstancy is evident during 1971–1975. That it is not perfect is less than surprising, as the data span a century during which financial institutions altered dramatically: witness the growth of building societies and, after 1970, the introduction of Competition and Credit Control regulations and of floating exchange rates. Even so, the evidence suggests that substantial benefits are available in practice from a progressive research strategy exploiting tests in both model design and model evaluation.

DATA APPENDIX

Legend

D_1	A dummy variable for World War I (= 1 for 1914–1918 inclusive, 0 elsewhere)
D_2	A dummy variable for 1921–1955, paralleling \bar{S} for phase-average data (= 1 for 1921–1955 inclusive, 0 elsewhere)
D_3	A dummy variable for World War II (= 1 for 1939–1945 inclusive, 0 elsewhere)
D_4	A dummy variable for Competition and Credit Control regulations (= 1 for 1971–1975, 0 elsewhere)
$G(\bar{p} + \bar{i})$	Growth rate of phase-average nominal income (fraction)
H	High-powered money (million £)
I	Real net national product (million 1929 £)
M	Money stock (million £)
N	Population (millions)
P	Deflator of I (1929 = 1.00)
P^*	Deflator of net national product in the United States (1929 = 1.00)
RL	Long-term interest rate (fraction)
RN	$RS \cdot (H/M)$
RS	Short-term interest rate (fraction)
\bar{S}	A dummy variable for phase observations 16–28 (1921–1955; = 1 for observations 16–28 inclusive, 0 elsewhere)
\bar{W}	A dummy variable for phase observations 13–15 and 26–28 (1918–1921 and 1946–1955; = -4, -3, -2, 8, 5, and 3 for phase observations 13, 14, 15, 26, 27, and 28, respectively; 0 elsewhere)

Coefficients and estimated standard errors of the dummies D_1 , D_2 , D_3 , (but not D_4), \bar{S} , and \bar{W} are reported times 100 for readability.

The data are as in Friedman and Schwartz (1982 tables 4.8, 4.9), but relevant series are rescaled proportionately from 1871 to 1920 to remove the break in 1920 when southern Ireland ceased to be part of the United Kingdom. Also, P , P^* , $G(\bar{p} + \bar{i})$, RS, and RL have been divided by 100 so that the values of P and

P^* in 1929 equal 1.00 (rather than 100) and the interest rates and $G(\bar{p} + \bar{i})$ are expressed as fractions (rather than as percentages).

Data Measurement

We record several important caveats about Friedman and Schwartz's data, over'apping those which they carefully document.

- (i) The choice of monetary measure seems too broad to represent transactions demand, yet too narrow for an overall index of "liquidity." M (above) is based on the U.K. monetary measure M_2 (pp. 111-4) and hence excludes interest-bearing liabilities of building societies but includes interest-bearing bank deposits. However, the building society movement has grown rapidly over the last century to become roughly equal in size to the whole commercial banking sector, and building society liabilities are among the most liquid assets available to the personal sector. Note that money-stock figures are centered on mid-years by averaging successive end-of-year values.
- (ii) The measurement of the price series (P) is adjusted for rationing and controls (pp. 115-20), with real income then derived by deflating nominal income. However, it is unreasonable to hold measured nominal income constant when measured prices are believed to be incorrect. Furthermore, over a century, one must be concerned about the effects on the measurement of P of the many dramatic changes that have occurred in quality-adjusted relative prices.
- (iii) Friedman and Schwartz emphasize an "errors-in-variables" paradigm and claim that the importance of errors in variables is reduced by phase-averaging but enhanced by differencing (p. 86). All the variables undoubtedly contain substantial measurement errors, especially when interpreted as correspondences to economically meaningful latent constructs, but time series of over a century will also contain systematic errors. Differencing would remove most effects of such errors, whereas averaging is of little help if errors are persistent (e.g., highly autoregressive).

Resolution of these difficulties is outside the scope of this paper, but they must influence the interpretation of the empirical evidence.

REFERENCES

- Artis, Michael J., "Book Review," *Economica*, May 1984, 51, 205-7.
- Barndorff-Nielsen, Ole, *Information and Exponential Families in Statistical Theory*, Chichester: Wiley, 1978.
- Barro, Robert J., *Macroeconomics*, 2nd ed., New York: Wiley, 1987.
- Boland, Lawrence A., *The Foundations of Economic Method*, London: Allen and Unwin, 1982.
- Box, George E. P. and Pierce, David A., "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models," *Journal of the American Statistical Association*, December 1970, 65, 1509-26.
- Bronfenbrenner, Jean, "Sources and Size of Least-Squares Bias in a Two-Equation Model," in William C. Hood and Tjalling C. Koopmans, eds., *Studies in Econometric Method*, New Haven: Yale University Press, 1953, 221-35.
- Brown, R. L., Durbin, James and Evans, J. M., "Techniques for Testing the Constancy of Regression Relationships over Time," *Journal of the Royal Statistical Society, Series B*, 1975, 37 (2), 149-92.
- Burns, Arthur F. and Mitchell, Wesley C., *Measuring Business Cycles*, New York: National Bureau of Economic Research, 1946.
- Campos, Julia and Ericsson, Neil R., "Econometric Modeling of Consumers' Expenditure in Venezuela," International Finance Discussion Paper No. 325, Board of Governors of the Federal Reserve System, Washington, DC, June 1988.
- _____, and Hendry, David F., "An Analogue Model of Phase-Averaging Procedures," *Journal of Econometrics*, March 1990, 43, 275-92.
- Chalmers, Alan F., *What Is This Thing Called Science? An Assessment of the Nature and Status of Science and Its Methods*, St. Lucia, Queensland: University of Queensland Press, 1976.
- Chow, Gregory C., "Tests of Equality between Sets of Coefficients in Two Linear Regressions," *Econometrica*, July 1960, 28, 591-605.
- Congdon, Tim, "Has Friedman Got It Wrong?" *The Banker*, July 1983, 117-25.
- Cuthbertson, Keith, "The Demand for M_1 : A Forward Looking Buffer Stock Model," *Oxford Economic Papers*, March 1988, 40, 110-31.
- Davidson, James E. H., "Disequilibrium Money: Some Further Results with a Monetary Model of the UK," in C. A. E. Good-

- hart, D. Currie, and D. T. Llewellyn, eds., *The Operation and Regulation of Financial Markets*, London: Macmillan, 1987, 125-49.
- _____, Hendry, David F., Srba, Frank and Yeo, Stephen, "Econometric Modelling of the Aggregate Time-Series Relationship between Consumers' Expenditure and Income in the United Kingdom," *Economic Journal*, December 1978, 88, 661-92.
- Desai, Meghnad, *Testing Monetarism*, London: Frances Pinter, 1981.
- Dickey, David A. and Fuller, Wayne A., "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, June 1979, 74, 427-31.
- _____, and _____, "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root," *Econometrica*, July 1981, 49, 1057-72.
- Dolado, Juan J. and Jenkinson, Tim, "Cointegration: A Survey of Recent Developments," Applied Economics Discussion Paper No. 39, Institute of Economics and Statistics, University of Oxford, November 1987.
- Dufour, Jean-Marie, "Recursive Stability Analysis of Linear Regression Relationships: An Exploratory Methodology," *Journal of Econometrics*, May 1982, 19, 31-76.
- Durbin, James and Watson, Geoffrey S., "Testing for Serial Correlation in Least Squares Regression. I," *Biometrika*, December 1950, 37, 409-28.
- _____, and _____, "Testing for Serial Correlation in Least Squares Regression. II," *Biometrika*, June 1951, 38, 159-78.
- Engle, Robert F., "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, July 1982, 50, 987-1007.
- _____, "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics," in Z. Griliches and M. D. Intriligator, eds., *Handbook of Econometrics*, Vol. 2, Amsterdam: North-Holland, 1984, 775-826.
- _____, "On the Theory of Cointegrated Economic Time Series, 1987," mimeo, University of California, San Diego, August 1987.
- _____, and Granger, Clive W. J., "Co-Integration and Error Correction: Representation, Estimation, and Testing," *Econometrica*, March 1987, 55, 251-76.
- _____, and Hendry, David F., "Testing Super Exogeneity and Invariance," Discussion Paper No. 89-51, University of California, San Diego, November 1989.
- _____, _____, and Richard, Jean-François, "Exogeneity," *Econometrica*, March 1983, 51, 277-304.
- Ericsson, Neil R. and Hendry, David F., "Conditional Econometric Modeling: An Application to New House Prices in the United Kingdom," in A. C. Atkinson and S. E. Fienberg, eds., *A Celebration of Statistics: The ISI Centenary Volume*, New York: Springer-Verlag, 1985, 251-85.
- Escribano, Alvaro, "Non-Linear Error-Correction: The Case of Money Demand in the U.K. (1878-1970)," mimeo, University of California, San Diego, December 1985.
- Fisher, R. A., "The Goodness of Fit of Regression Formulae, and the Distribution of Regression Coefficients," *Journal of the Royal Statistical Society*, 1922, 85 (4), 597-612.
- Florens, Jean-Pierre and Mouchart, Michel, (1985a) "Conditioning in Dynamic Models," *Journal of Time Series Analysis*, 1985, 6 (1), 15-34.
- _____, and _____, (1985b) "A Linear Theory for Noncausality," *Econometrica*, January 1985, 53, 157-75.
- Friedman, Milton, *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.
- _____, "The Quantity Theory of Money—A Restatement," in M. Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago: University of Chicago Press, 1956, 3-21.
- _____, and Schwartz, Anna J., *Monetary Trends in the United States and the United Kingdom: Their Relation to Income, Prices, and Interest Rates, 1867-1975*, Chicago: University of Chicago Press, 1982.
- Gilbert, Christopher L., "Professor Hendry's Econometric Methodology," *Oxford Bul-*

- letin of Economics and Statistics*, August 1986, 48, 283–307.
- Godfrey, Leslie G., "Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables," *Econometrica*, November 1978, 46, 1293–1301.
- Goodhart, Charles A. E., "Monetary Trends in the United States and the United Kingdom: A British Review," *Journal of Economic Literature*, December 1982, 20, 1540–51.
- _____, *Monetary Theory and Practice: The UK Experience*, London: Macmillan, 1984.
- Granger, Clive W. J., "Some Properties of Time Series Data and Their Use in Econometric Model Specification," *Journal of Econometrics*, May 1981, 16, 121–30.
- Hallman, Jeff, "Attractor Sets: A Nonlinear Extension of the Cointegration Concept," mimeo, University of California, San Diego, November 1987.
- Harvey, Andrew C., *The Econometric Analysis of Time Series*, Oxford: Philip Allan, 1981.
- Hawtrey, R. G., *A Century of Bank Rate*, London: Longmans, Green, 1938.
- Hendry, David F., "Predictive Failure and Econometric Modelling in Macroeconomics: The Transactions Demand for Money," in P. Ormerod, ed., *Economic Modelling*, London: Heinemann, 1979, 217–42.
- _____, "Econometric Modelling: The 'Consumption Function' in Retrospect," *Scottish Journal of Political Economy*, November 1983, 30, 193–220.
- _____, ed., *Economic Modelling with Cointegrated Variables* (special issue), *Oxford Bulletin of Economics and Statistics*, August 1986, 48.
- _____, "Econometric Methodology: A Personal Perspective," in T. F. Bewley, ed., *Advances in Econometrics*, Vol. 2, Cambridge: Cambridge University Press, 1987, 29–48.
- _____, "The Encompassing Implications of Feedback versus Feedforward Mechanisms in Econometrics," *Oxford Economic Papers*, March 1988, 40, 132–49.
- _____, *PC-GIVE: An Interactive Econometric Modelling System*, Version 6.0/6.01, Oxford: Institute of Economics and Statistics and Nuffield College, University of Oxford, 1989.
- _____, and Ericsson, Neil R., "Assertion without Empirical Basis: An Econometric Appraisal of 'Monetary Trends in...the United Kingdom' by Milton Friedman and Anna Schwartz," in Bank of England Panel of Academic Consultants, *Monetary Trends in the United Kingdom*, Panel Paper No. 22, October 1983, 45–101 (with additional references).
- _____, and _____, "Assertion without Empirical Basis: An Econometric Appraisal of *Monetary Trends in...the United Kingdom* by Milton Friedman and Anna J. Schwartz," International Finance Discussion Paper No. 270, Board of Governors of the Federal Reserve System, Washington, DC, December 1985.
- _____, and _____, "Prolegomenon to a Reconstruction: Further Econometric Appraisal of *Monetary Trends in...the United Kingdom* by Milton Friedman and Anna J. Schwartz," mimeo, Board of Governors of the Federal Reserve System, Washington, DC, March 1986.
- _____, and Mizon, Grayham E., "Serial Correlation as a Convenient Simplification, Not a Nuisance: A Comment on a Study of the Demand for Money by the Bank of England," *Economic Journal*, September 1978, 88, 549–63.
- _____, and _____, "Procrustean Econometrics: Or Stretching and Squeezing Data," in Clive W. J. Granger, ed., *Modelling Economic Series: Readings in Econometric Methodology*, Oxford: Clarendon Press, 1990, 121–36.
- _____, and Neale, Adrian J., "Monte Carlo Experimentation Using PC-NAIVE," in G. F. Rhodes, Jr., ed., *Advances in Econometrics*, Vol. 6, Greenwich, CT: JAI Press, 1987, 91–125.
- _____, _____, and Srba, Frank, "Econometric Analysis of Small Linear Systems Using PC-FIML," *Journal of Econometrics*, May/June 1988, 38, 203–26.
- _____, Pagan, Adrian R. and Sargan, J. Denis, "Dynamic Specification," in Z. Griliches and M. D. Intriligator, eds., *Handbook of Econometrics*, Vol. 2, Amsterdam: North-Holland, 1984, 1023–1100.

- _____ and Richard, Jean-François, "On the Formulation of Empirical Models in Dynamic Econometrics," *Journal of Econometrics*, October 1982, 20, 3-33.
- _____ and _____, "The Econometric Analysis of Economic Time Series," *International Statistical Review*, August 1983, 51, 111-63.
- _____ and _____, "Recent Developments in the Theory of Encompassing," in Bernard Cornet and Henry Tulkens, eds., *Contributions to Operations Research and Economics: The Twentieth Anniversary of CORE*, Cambridge, MA: MIT Press, 1989, 393-440.
- _____ and Wallis, Kenneth F., eds., *Econometrics and Quantitative Economics*, Oxford: Blackwell, 1984.
- Herschel, J., *A Preliminary Discourse on the Study of Natural Philosophy*, London: Longman, Rees, Brown & Green and John Taylor, 1830.
- Jarque, Carlos M. and Bera, Anil K., "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals," *Economics Letters*, 1980, 6 (3), 255-9.
- Johnston, J., *Econometric Methods*, New York: McGraw-Hill, 1963.
- Judd, John P. and Scadding, John L., "The Search for a Stable Money Demand Function: A Survey of the Post-1973 Literature," *Journal of Economic Literature*, September 1982, 20, 993-1023.
- Kiviet, Jan F., "Model Selection Test Procedures in a Single Linear Equation of a Dynamic Simultaneous System and Their Defects in Small Samples," *Journal of Econometrics*, June 1985, 28, 327-62.
- _____, *Testing Linear Econometric Models*, Amsterdam: Amsterdam University Press, 1987.
- Klovland, Jan T., "The Demand for Money in the United Kingdom, 1875-1913," *Oxford Bulletin of Economics and Statistics*, August 1987, 49, 251-71.
- Koopmans, Tjalling C., "When Is an Equation System Complete for Statistical Purposes?" in Tjalling C. Koopmans, ed., *Statistical Inference in Dynamic Economic Models*, New York: Wiley, 1950, 393-409.
- Lakatos, Imre, "Falsification and the Methodology of Scientific Research Programmes," in I. Lakatos and A. Musgrave, eds., *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, 1970, 91-196.
- Leamer, Edward E., *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, New York: Wiley, 1978.
- _____, "Let's Take the Con Out of Econometrics," *American Economic Review*, March 1983, 73, 31-43.
- Longbottom, Andrew and Holly, Sean, "Econometric Methodology and Monetarism: Professor Friedman and Professor Hendry on the Demand for Money," Discussion Paper No. 131, London Business School, February 1985.
- Lubrano, Michel, Pierse, Richard G. and Richard, Jean-François, "Stability of a U.K. Money Demand Equation: A Bayesian Approach to Testing Exogeneity," *Review of Economic Studies*, August 1986, 53, 603-34.
- MacKinnon, James G., "Model Specification Tests Against Non-Nested Alternatives," *Econometric Reviews*, 1983, 2 (1), 85-158.
- _____ and White, Halbert, "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, September 1985, 29, 305-25.
- Mayer, Thomas, "Monetary Trends in the United States and the United Kingdom: A Review Article," *Journal of Economic Literature*, December 1982, 20, 1528-39.
- McAleer, Michael, Pagan, Adrian R. and Volker, Paul A., "What Will Take the Con Out of Econometrics?" *American Economic Review*, June 1985, 75, 293-307.
- Milbourne, Ross, "Optimal Money Holding Under Uncertainty," *International Economic Review*, October 1983, 24, 685-98.
- Miller, Merton H. and Orr, Daniel, "A Model of the Demand for Money by Firms," *Quarterly Journal of Economics*, August 1966, 80, 413-35.
- Mizon, Grayham E., "The Encompassing Approach in Econometrics," in D. F. Hendry and K. F. Wallis, eds., *Econometrics and Quantitative Economics*, Oxford: Blackwell, 1984, 135-72.
- _____ and Richard, Jean-François, "The En-

- compassing Principle and Its Application to Testing Non-Nested Hypotheses," *Econometrica*, May 1986, 54, 657-78.
- Moore, Basil J., "Monetary Trends in the United States and in the United Kingdom, A Review," *The Financial Review*, May 1983, 18, 146-66.
- Nicholls, Desmond F. and Pagan, Adrian R., "Heteroscedasticity in Models with Lagged Dependent Variables," *Econometrica*, July 1983, 51, 1233-42.
- Nickell, Stephen, "Error Correction, Partial Adjustment and All That: An Expository Note," *Oxford Bulletin of Economics and Statistics*, May 1985, 47, 119-29.
- Patterson, Kerry D., "The Stability of Some Annual Consumption Functions," *Oxford Economic Papers*, March 1986, 38, 1-30.
- Pesaran, M. Hashem, "Comparison of Local Power of Alternative Tests of Non-Nested Regression Models," *Econometrica*, September 1982, 50, 1287-1305.
- Phillips, Peter C. B., "Time Series Regression with a Unit Root," *Econometrica*, March 1987, 55, 277-301.
- _____, "Reflections on Econometric Methodology," *Economic Record*, December 1988, 64, 344-59.
- _____, and Hansen, Bruce E., "Statistical Inference in Instrumental Variables Regression with I(1) Processes," *Review of Economic Studies*, January 1990, 57, 99-125.
- Popper, Karl R., *The Logic of Scientific Discovery*, London: Hutchinson, 1959.
- Ramsey, J. B., "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis," *Journal of the Royal Statistical Society, Series B*, 1969, 31 (2), 350-71.
- Salmon, Mark, "Error Correction Mechanisms," *Economic Journal*, September 1982, 92, 615-29.
- Sargan, J. Denis, "The Estimation of Economic Relationships Using Instrumental Variables," *Econometrica*, October 1958, 26, 393-415.
- _____, "Wages and Prices in the United Kingdom: A Study in Econometric Methodology," in P. E. Hart, G. Mills and J. K. Whitaker, eds., *Econometric Analysis for National Economic Planning*, Colston Papers, Vol. 16, London: Butterworths, 1964, 25-63; reprinted in D. F. Hendry and K. F. Wallis, eds., *Econometrics and Quantitative Economics*, Oxford: Blackwell, 1984, 275-314.
- _____, (1980a) "Some Approximations to the Distribution of Econometric Criteria which Are Asymptotically Distributed as Chi-Squared," *Econometrica*, July 1980, 48, 1107-38.
- _____, (1980b) "Some Tests of Dynamic Specification for a Single Equation," *Econometrica*, May 1980, 48, 879-97.
- _____, and Bhargava, Alok, "Testing Residuals from Least Squares Regression for Being Generated by the Gaussian Random Walk," *Econometrica*, January 1983, 51, 153-74.
- Sims, Christopher A., "Macroeconomics and Reality," *Econometrica*, January 1980, 48, 1-48.
- Smith, Gregor W., "A Dynamic Baumol-Tobin Model of Money Demand," *Review of Economic Studies*, July 1986, 53, 465-9.
- Spanos, Aris, *Statistical Foundations of Econometric Modelling*, Cambridge: Cambridge University Press, 1986.
- Starr, Ross M., "Variation in the Maturity Composition of Debt and Behavior of the Monetary Aggregates: The Maturity Shift Hypothesis," *Economic Review* (San Francisco), conference supplement, Fall 1982, 202-33.
- Stock, James H., "Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors," *Econometrica*, September 1987, 55, 1035-56.
- Trundle, John M., "The Demand for M1 in the UK," mimeo, Bank of England, 1982.
- White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, 48, 817-38.
- _____, *Specification Analysis in Econometrics*, manuscript, 1988; Cambridge: Cambridge University Press, forthcoming.

Alternative Approaches to Analyzing Economic Data

By MILTON FRIEDMAN AND ANNA J. SCHWARTZ*

This paper compares two approaches to the analysis of economic data: the exclusively econometric approach recommended by Hendry and Ericsson and the more eclectic approach that we employed in our books Monetary History and Monetary Trends. It concludes that the article by Hendry and Ericsson is mislabeled. Their article is not in any relevant sense an evaluation of our "empirical model of U.K. money demand." Rather, it uses one equation from our book as a peg on which to hang an exposition of sophisticated econometric techniques. Insofar as their empirical findings do bear on ours, they simply confirm some of our principal results and contradict none. (JEL 036, 212)

In response to the "on the one hand" and "on the other hand" economist, Frank H. Knight's favorite remark was, "Yes, indeed, there are two sides to every issue: the right side and the wrong side." David Hendry and Neil Ericsson (1991) (hereafter HE) take the same position—though in their case seriously not jocularly—about how to do empirical research in economics: HE's way and the wrong way. We are more eclectic. We believe that there is no magic formula for wringing reasonable conjectures from refractory and inaccurate evidence. The HE approach is one way: start with a collection of numerical data bearing on the question under study, subject them to sophisticated econometric techniques, place great reliance on tests of significance, and end with a single hypothesis (equation), however complex, supposedly "encompassing"—to use one of HE's favorite terms—all subhypotheses. Another way is to examine a wide variety of evidence, quantitative and nonquantitative, bearing on the question under study; test results from one body of evidence on other bodies, using econometric techniques as one tool in this process, and build up a collection of simple hypotheses that may or may not be readily viewed as components of a broader all-embracing

hypothesis; and, finally, test hypotheses on bodies of data other than those from which they were derived. Both ways—and no doubt still others as well—have their use. None, we believe, can be relied on exclusively.

I. A Comparison of Approaches

The wide difference between HE's approach and ours is clear from HE's section II, in which they "discuss the money-demand relationships they [we] estimated from the phase-average data for the United Kingdom" (HE, p. 9). From the literally hundreds of regressions in *Monetary Trends* (Friedman and Schwartz, 1982), they included four in their table 1 and concentrated their fire on only one of them. They assert that our regression for U.K. money demand is misspecified, "is not an adequate characterization of the data" (HE, p. 15) and does not support any of our "main claims about the United Kingdom" (HE, p. 32). They use that regression as a textbook example to illustrate a collection of sophisticated statistical techniques and to call the reader's attention to the technical statistical literature dealing with those techniques—most of it published after our book went to press.¹ That may be a good way to intro-

*Friedman is Senior Research Fellow, Hoover Institution, Stanford University, and Professor Emeritus of Economics, University of Chicago; Schwartz is Research Associate, National Bureau of Economic Research.

¹Out of 108 total references in HE's list, 70 were published after 1981, the year our book went to press; also, 20 out of 23 articles that were authored by Hendry alone or with collaborators.

duce students to the particular techniques covered and to the associated literature—which, we hasten to add, we are not competent to deploy, let alone judge—but it provides little if any evidence that is relevant to judging the validity of our analysis of money demand.

Hendry and Ericsson are entirely right that the one regression they subject to detailed scrutiny is not *by itself* adequate support for our “main claims about the United Kingdom” (see the Addendum to this article, in which Milton Friedman relates an experience that contributed to his skepticism about relying on a single regression produced by the HE approach). If that regression had been the only source of our “claims,” we would never have made them. As we emphasize in our book—partly in a passage that HE quote but also elsewhere—we do not have much confidence in “what has become the prevailing fashion in econometric work, the immediate computation of multiple regressions including all variables that can reasonably be regarded as relevant” (p. 215). The regression they analyze was selected from a table on page 282 (all unidentified page references herein are to Friedman and Schwartz [1982]) that contains six regressions, three for the United States and three for the United Kingdom, and was itself a prelude to presenting two equations “as a final summary of our results” for the United States and the United Kingdom combined (p. 284).

By HE’s standards, the prior 281 pages of our book are mostly worthless. Indeed, as we shall shortly demonstrate, while they may have skimmed those pages, they have apparently not thought it worthwhile to read them carefully. Those pages were not devoted, à la HE, to “representing the joint density of [a limited set of variables] in terms of an autoregressive-distributed lag model,” then proceeding to simplify “[t]he conditional model...to an ECM [error-correction model],” and to evaluating it “in light of the model design criteria” listed in their table 2 (HE, pp. 22–3). Instead, the first 204 of those 281 pages present our theoretical framework, our statistical framework, the basic data, and an overview of the

movements of money, income, and prices over the century our data cover. We end our broad overview by noting “three phenomena that require further interpretation: first, the common movement of nominal money and nominal income; second, the largely common movement in velocity in the United States and in the United Kingdom; third, the rather different relations in the two countries between movements in money, on the one hand, and real income and prices on the other” (pp. 184–5). Only then, in Chapter 6, “Velocity and the Demand for Money,” do we explore the first of the three phenomena, and, in subsequent chapters, the other two phenomena.

Throughout, our aim is to explain movements longer than a business cycle and to abstract from strictly intracyclical effects. That aim is both less and more ambitious than HE’s. It is less ambitious in that we limit attention to secular effects, while they seek a single econometric specification that simultaneously describes cyclical and secular movements. It is more ambitious in that we use as wide a range of evidence as we can, including qualitative examination of historical experience as well as numerical data for more than a century for both the United States and the United Kingdom. A major purpose of our book was to “give numerical content to some of the elements and hypotheses embedded in” “the broad theoretical framework that guided our earlier study of United States monetary history” “as well as some of the generalizations suggested by *A Monetary History*” (p. 13 *passim*). Accordingly, we frequently draw on evidence for the United States presented in the earlier book (Friedman and Schwartz, 1963). In particular, we use independent data for the United Kingdom to test generalizations initially suggested by U.S. experience.

The difference between HE’s approach and ours is exemplified by one of the test “criteria” by which they judge our equation unacceptable, namely that “[t]he values of $\hat{\sigma}$ [the standard error of estimate] for the corresponding subperiods [the first and second half, 1875–1928 and 1928–1975] are markedly different: 2.8 percent and 6.0 per-

cent, respectively" (HE, p. 13). On page 175 of our book, we noted that "For the United Kingdom, for the period before 1914, the extensive use of interpolation . . . biases sharply downward both the initial and the residual variation." In consequence, if the error of estimate had been the same in the two periods HE compare, we would have regarded that fact as evidence *against*, not for, our equation. That would have meant that we had erroneously attributed pure measurement error to economic variables. Similarly, the finding by HE that the equations they prefer (their equations 10 and 11) have the same residual variability before and after 1914 (not 1928) is, in our view, evidence against, not for, their equations.

They say, we "do not formally test for constancy" (HE, p. 13) by which they mean we do not use the particular statistical tests that they advocate; but we certainly do test for constancy by comparing regressions for different periods, trying to isolate features of the statistical data that may bias comparisons, drawing on historical data, and in a variety of other ways. We regard such tests as far more reliable and, to use a statistical term, robust than the formal statistical tests used by HE.

Another example is their assertion that "price homogeneity" incorporated in our model is rejected by their statistical test and their accompanying comment that, when we "tested for trends and price homogeneity in more restrictive models, they [we] did not obtain rejections" (HE, p. 15). From our viewpoint, the tests for price *and* population homogeneity on pages 253-9 were less, not more, restrictive than HE's test.² Our tests were for both the United States and the United Kingdom, used both levels and rates of change, and allowed for the "regression effect," of which more later, as well as for a variety of possible explanatory variables. We concluded that the results were "reasonably consistent with the" assumption that "any change in prices or population implied an equal percentage change in the quantity of

money demanded (mathematically, the demand for money in nominal terms is homogeneous of the first degree in population and prices)" (pp. 258, 253). We hasten to add that our conclusion was for a long-run demand function, not for a function seeking to cover intracyclical effects.

Still another example is HE's rejection of our use of phase averages and their decision to analyze our "original annual observations" (HE, p. 18). That is a correct decision for their objective: a single equation that "encompasses" both secular and cyclical effects. However, that was not our objective. We devoted pages 79-97 to testing whether the phase averages and rates of change contain a residual cyclical element and whether they are infected by a spurious serial correlation. We concluded "that our phase bases do eliminate the bulk of the systematic cyclical fluctuation" and "that the 'noise' introduced by the serial correlations is sufficiently small relative to the systematic variation we are trying to describe that it can for the most part be neglected" (pp. 81, 97). True enough, we discovered, as HE do, that annual observations give essentially the same results as phase-average data for the regression they analyze, and its counterpart for the United States (see Friedman, 1988 footnote 16). However, we doubt that we would have settled on those specific final equations if we had relied solely on annual data in our prior exploratory investigation. Throughout that investigation, we did not have to allow for cyclical effects, or explore a "(dynamic) error-correction representation of the relevant variables" (HE, p. 20), as we would have had to do if we had followed HE's prescription. Hence, we remain unpersuaded that our use of phase-average data was a mistake.

II. Exogeneity

Hendry and Ericsson put much emphasis on "concepts of exogeneity," listing four distinct concepts, and stating: "In no case is it legitimate to 'make variables exogenous' simply by not modeling them" (HE, p. 21). We do not sympathize with such commandments. In our view, exogeneity is not an

²So far as we can see, HE nowhere even try to test for population homogeneity.

invariant statistical characteristic of variables. Everything depends on the purpose. In economic analysis, it may be appropriate to regard a variable as exogenous for some purposes and as endogenous for others. A simple example is the quantity of money. For the United States after World War I, we believe it is appropriate to regard the money stock as exogenous (i.e., determined by the monetary authorities) in an economic analysis of long-run money demand. We do not believe it would be equally appropriate to do so for week-to-week or month-to-month movements for which, in HE's words, "the money stock appears to be endogenously determined by the decisions of the private sector" (HE, p. 27). For the period before World War I, as we repeatedly state in both *Monetary History* and *Monetary Trends*, it is not appropriate to regard the money stock as exogenous even for money-demand analysis, particularly for phase averages. For that period, the money stock is best regarded as endogenous, because the United States and much of the rest of the world was on a gold standard. Even for the post-World War I period, it would not be appropriate to regard the money stock as exogenous in a broader study of the Federal Reserve as a political institution, created by Congress, subject to its ultimate control, with members of its Board appointed by the President. For such studies, of which there are a good number, the money stock is best regarded as endogenous.

After their own examination of "exogeneity" and "super exogeneity" for one of their regressions, HE assert that "This evidence is inconsistent with the hypothesis that, over the period 1878–1970, exogenous money determined prices in the United Kingdom via a stable demand function, *precisely because we have established a constant money-demand model conditional on prices*" (HE, p. 30). We interpret their italics as implying that we asserted the hypothesis that they regard as contradicted. We clearly did nothing of the kind. In particular, a large part of our next chapter (Ch. 7) explores the direction of influence of the variables HE consider as well as the direction of influence between the United States and the United

Kingdom. We distinguish different monetary regimes and examine separately the gold-standard period and the variable-exchange-rate period. At the outset of Chapter 7, we note that "one possible explanation for the similar behavior of [velocity] is that the two countries were part of a single economic entity" (p. 305). We go on to note that before World War I, when both countries were on a gold or sterling standard, "as we demonstrated in *A Monetary History*, ... there was a good deal of leeway for domestic monetary policy over short periods, but over periods of more than a few years, the quantity of money in each country was determined by the requirement that the price levels of the two countries move roughly in step in order to preserve equilibrium in the balance of payments" (p. 306). Clearly, we are not guilty of the sin that we suspect, perhaps wrongly, HE accuse us of.

III. The Regression Effect

On the purely statistical level, we tried throughout to avoid making arbitrary assumptions about exogeneity. For example, in the first table in Chapter 6 (table 6.1, p. 212), we label one set of equations "Level of Money: Income Assumed Exogenous" and another "Level of Income: Money Assumed Exogenous" and do the same for rate-of-change calculations. Similar caution in later tables derives in large part from our trying systematically to allow for what we referred to earlier as "the regression effect."

The regression effect refers to the bias introduced by stochastic disturbances ("error") affecting a variable treated as exogenous. Indeed, regression analysis got its name from this bias.³ Put formally, consider a relation between two variables (x and y) that is strictly linear. Given errors of mea-

³In studying the relation between the heights of fathers and sons, Sir Francis Galton discovered that the average height of sons of tall (short) fathers "regressed" to the mean (i.e., differed from the average height of all sons by less than their fathers' height differed from the average height of all fathers) and, simultaneously, the average height of the fathers of tall (short) sons also "regressed to the mean."

surement, a sample of observed values of the two variables will yield a bivariate scatter, rather than a strictly linear relation. The calculated regression of y on x will be flatter than the "true" relation, and that of x on y will be steeper. Harold Hotelling pointed out in 1933 how potent a source of economic fallacies the regression effect can be if "errors of measurement" are interpreted to include all stochastic disturbances affecting the variables under study (Hotelling, 1933, 1934; Horace Secrist, 1934; Friedman and Simon Kuznets, 1945 pp. 325–38; Friedman, 1957 Ch. 3).

A major difference between our statistical approach and HE's is that we allow systematically for the regression effect. HE obliquely recognize this difference in their data appendix, where they note near the end that we "emphasize an 'errors-in-variables' paradigm" (HE, p. 34). Nowhere else do they mention or refer to the regression effect or seek to allow for it. In our opinion, it is often a more important source of error—because it introduces systematic bias—than is the residual error described by the standard error of estimate to which HE give exclusive attention.

In our analysis, as in many economic analyses, the so-called "independent" or "exogenous" or "predetermined" variables are infected by error in a double sense. In the first place, they are not always precise empirical counterparts to the economic variables that theory suggests including. To illustrate, we regard the yield on physical assets as one of the interest rates that can be expected to affect the real quantity of money demanded. However, despite considerable exploration, we could not find a direct empirical counterpart. We finally settled on using the rate of change of nominal income as a "proxy for the nominal return on physical assets" (pp. 274–80). As a second example, we consider as another relevant variable, the own yield on money, to which HE refer in the final paragraph of their section IV, citing a 1987 article, but do not empirically explore. Here again, we were unable to find a reliable reasonably direct measure, though further statistical exploration of basic data may enable such a mea-

sure to be constructed. Instead, we use an indirect measure suggested by Benjamin Klein (1970) and described by HE in their footnote 26. (We examine the whole issue on pp. 259–73.) In both cases, even if the proxy were measured precisely, there would remain measurement error equal to the difference between the relevant economic variable and the proxy used as its empirical counterpart.

In the second place, the proxy variable is never measured precisely, so there is a measurement error in the ordinary sense that adds to the total measurement error. When these measured variables are used in a regression as independent variables, their coefficients are biased estimates of the underlying theoretical coefficients. One way to estimate the possible size of the bias is to reverse the direction of regression: in the bivariate case, to estimate both a y -on- x regression and an x -on- y regression. The resulting estimates of the theoretical coefficients provide an upper and a lower limit. A similar procedure can be used to get upper and lower limits in multiple regressions. We discussed this issue at some length and dealt with it by consistently calculating such limits for essentially all of the parameters we estimated.⁴ In principle, the limits should be made still wider to allow for the usual standard error of estimate of the parameters arising from the stochastic element in the dependent variable. We did not do so; however, we consistently presented t ratios, from which what is generally called the "sampling error" can be readily calculated, so that any interested reader can combine the two.

An example shows the importance of the regression effect compared with the conventional "sampling error" that is HE's sole concern. Consider the coefficient of the interest rate that HE refer to in their section

⁴See especially Chapter 5, footnotes 28 and 29, pp. 173–4; Chapter 6, footnote 7, p. 211, footnote b to table 6.1, p. 213, footnote 18, pp. 224–5; and the text to which these footnotes are attached, as well as comments scattered throughout the text in discussing particular sets of results.

II and plot in their figure 5. Our upper and lower point estimates allowing for the regression effect differ by 28.6 (table 6.15, p. 285). One estimate of the standard error of the coefficient is 3.26.⁵ Even four times the standard error (plus or minus twice the standard error) is 13.04, or less than half the regression effect. Combining the sampling error and the regression effect gives a range for the coefficient from -4.6 to -46.3 , the basis for our statement that this coefficient was "much less precisely estimated" than the income elasticity, for which the regression effect was 0.08, the standard error of estimate of the parameter 0.005, so the combined range was 0.87–0.97.

This example is, we believe, fairly representative with respect to the relative importance of the regression effect and the regularly computed standard error of estimate of parameters. It suggests why we attribute so much importance to the regression effect, or alternatively, why we regard the "errors-in-variable model" as appropriate for economic data that are subject to large differences between the theoretical variables and their statistical counterparts and to large measurement errors.⁶

IV. The Proof of the Pudding

Hendry and Ericsson's final products which, by comparison with our results, they describe as a "better-fitting, constant, dynamic, error-correction (cointegration) model" (HE, p. 8) and which satisfy all of

their statistical "criteria for evaluating and designing models" (HE, p. 19) are their regressions 10 and 11, the first, with 10 parameters, omitting the five final years, 1971–1975; the second, with 12 parameters, covering the whole period, 1878–1975. These are in turn simplifications of the "general autoregressive-distributed lag" regression in their table 3 containing 37 parameters and omitting the final five years. How do their results compare with ours? In terms of the long-run effects we were interested in, they simply confirm a few of our results.

A. Income Elasticity of Money Demand

Our point estimate of the elasticity of per capita real demand is 0.88; allowing for sampling error by adding plus or minus twice the standard error of estimate of the relevant parameter gives a range of 0.87–0.89; allowing also for the regression effect gives, as noted earlier, a range of 0.87–0.97. The regression yielding these estimates has six parameters.

Neither HE's regression 10 nor their regression 11 contains an explicit income term. The reason is presumably their prior conclusion, embodied in regression 9 and apparently regarded by them as confirmed by the 37-parameter regression in their table 3, that the income elasticity of aggregate money demand is unity. Nowhere do they allow for population (unless implicitly in the constant terms of regressions 10 and 11). Since we decided that the elasticity of money demand with respect to population could be taken as unity, the aggregate elasticity can be expected to be closer to unity than the per capita elasticity. The regression in HE's table 3 does include income separately and so permits an elasticity other than one. If we convert it into a long-run regression by adding the coefficients of the separate lag terms, as is done in the final column of the table, the result is an eight-parameter regression and a point estimate of the income elasticity of 1.1. Clearly there is no contradiction between their estimate and ours if allowance is made for both sampling error and the regression effect.

⁵"One estimate" because it is from one of the two regressions from which we calculate the limits (e.g., y on x). The other regression (e.g., x on y) would give a somewhat different estimate.

⁶In re measurement error, in a later chapter (Ch. 8), we use a highly indirect approach to estimate the pure measurement error in our data. For the phase-average level of nominal income, we estimate that the standard deviation of pure measurement error is 2.5 percent for the United States, 4.5 percent for the United Kingdom. For the rate of change of nominal income computed from the phase averages, we estimate the corresponding standard deviations to be 0.74 of a percentage point for the United States and 0.58 of a percentage point for the United Kingdom (pp. 350–1). These are appreciable, yet not unreasonable.

B. Interest Semielasticity of Money Demand

Our point estimate of the semielasticity of money demand with respect to the differential yield on money (the short-term interest rate minus our proxy for the own-rate on money) is -11.16 . This implies that a *one percentage point* increase in the differential interest rate reduces the real quantity of money demanded per capita by 11.16 percent. As noted earlier, we regard this estimate as not very well determined and the range, including both sampling and regression effects, is extremely wide, from -4.6 to -46.3 , though, as theory would lead one to expect, negative throughout.

HE produce no strictly comparable elasticity. They use the short-term interest rate but do not allow for the own rate.⁷ In commenting on their equation 9, they state that it "implies that a one-percentage-point increase in the short-term interest rate... reduces [velocity]... by 7 percent in the long run," (p. 25) or that their estimate of the semielasticity of the short-term interest rate is -7 , well within our estimated range.⁸

C. Postwar Readjustment and Demand Shift

Our regression includes two dummy variables to allow for postwar readjustment and a demand shift. HE remark that "many investigators would regard the need for the data-based shift dummy \bar{S} spanning one-

third of the sample as *prima facie* evidence against the model's constancy" (HE, p. 13). However, the regression in their table 3 has three dummy variables, one of which is the precise counterpart of our demand shift dummy \bar{S} ; the other two, like our postwar adjustment dummy, allow for wartime effects. Their regression 10 combines the two wartime dummies but does not explicitly use a demand-shift dummy. Their regression 11 introduces an additional dummy for the 1971-1975 period.

The coefficient of our demand-shift dummy is 0.21, implying that the demand for money shifted upward by 21 percent during the period in question. Adding and subtracting two standard errors of estimate gives a range of 0.15-0.27. HE's table 3 gives a point estimate, in units comparable to our estimate, of 0.06 but with a very large standard error, so that adding and subtracting two standard errors gives a range from -0.29 to $+0.41$. Commenting on their "simplified" regression 10, HE note that it "exhibits multiple equilibria, with two corresponding to the long-run solution (9) and a third being that solution shifted by 20 percent. In that sense, the results are consistent with the use of an adjustment factor of about that order [i.e., 21 percent] by Friedman and Schwartz" (HE, p. 26). We have not been able to devise any ready way to compare our estimate of the postwar readjustment effect with the effect of their two wartime dummies.

D. Differences in Scope

So far, the HE results simply confirm ours. In addition, we 1) estimated the effect of the proxy yield on physical assets, a variable they do not consider; 2) allowed for the effect of the own-yield on money, which they postpone for future work; 3) compared demand functions in the United States and the United Kingdom, concluding that the only important difference was a higher income elasticity for the United States than for the United Kingdom; and 4) estimated two regressions for the two countries combined, including one additional dummy to allow for the difference in income elasticity.

⁷However, in their footnote 26, HE assert that they experimented with using our proxy and that "The resulting estimates are very similar to those found using" only the short-term rate.

⁸In one of our preliminary regressions that HE do not refer to, the point estimate of the semielasticity for the short-term rate alone is -3.0 ; the range, allowing for both sampling error and the regression effect, is from -1.4 to -11.1 , again including their -7 (see tables 6.12 and 6.N.3, pp. 272-3 and 277).

The long-run regression derived from HE's table 3 contains both a short-term and a long-term interest rate. However, as judged from the final column, the coefficient of neither comes close to being significantly different statistically from 0, and any range constructed from those point estimates is so wide as to be meaningless.

ties. HE, as part of their more ambitious agenda, estimate the short-run error-correction process, something that we explicitly excluded from our study.

E. Encompassing

Hendry and Ericsson put a great deal of emphasis on whether one "model" encompasses another. They assert that "a necessary condition for encompassing is *variance dominance*, where one equation variance-dominates another if the former has a smaller error variance" (HE, p. 22). A footnote attached to this sentence states: "Formally, variance dominance refers to the underlying (and unknown) error variances. Without loss of clarity, we often will say a model variance-dominates another if the *estimated residual* variance of the former is smaller than of the latter." In other words, they judge the validity of their hypotheses by the data from which they derive them!

In this spirit, they boast that "the error variance of (10) is less than one-tenth of that in" our regression (HE, p. 27); but this is to compare apples and oranges. Their regression 10 has a first difference or a rate of change as a dependent variable. The regression they compare it with has the logarithm of the level of money as the dependent variable. They present only two regressions in log levels: their regression 9 and the regression in their table 3. The standard error of estimate for regression 9 is nearly twice the standard error of estimate of our regression; for the regression in their table 3, it is nearly three times higher.⁹

We also estimated a parallel regression in rates of change, comparable to HE's regressions 10 and 11. It has six parameters and a standard error of estimate of 1.34, compared with 1.424 for their 10-parameter regression 10, which covers a shorter period than ours, and 1.478 for their 12-parameter

regression 11, which covers the same period as ours. In their terms, our regression variance-dominates theirs. They obliquely recognize the problem and try to justify their procedure: "Although (10) has a rate of change as the dependent variable, it is an equation in log-levels because of the error-correction term..., so direct comparison [with our log-level equation] is valid" (HE, p. 27). However, their "error-correction terms" are useful primarily for short-run analysis of levels, not for the long-run effects that were our sole concern.

HE's regressions 10 and 11 are misleading for a very different reason as well. The same variable, Δp_t , is included on both sides of the equation: as part of the dependent variable and also as an independent variable. Adding Δp_t to both sides of the regression does not change its statistical characteristics ($\hat{\sigma}$, the standard errors of the parameters, and the coefficients of all terms other than Δp_t are unchanged). What it does make clear is that the regression "explains" a change in the nominal quantity of money, not the real quantity of money (is it, perhaps, best interpreted as an estimate of a short-run supply curve of nominal money?).

We have been unable to duplicate regression 10 precisely, but we have come reasonably close. Omitting Δp_t from the right-hand side nearly doubles the standard error of estimate (2.83 percent versus 1.50 percent).¹⁰ A parallel equation with Δm_t as the independent variable, and omitting Δp_t , gives a standard error of estimate one-third larger than that from our estimate of regression 10 (2.07 percent versus 1.50 percent). That is a reasonably satisfactory empirical equation for the change in the nominal quantity of money. However, we are hard put to construct any satisfactory theoretical interpretation of the regression.

HE refer favorably to papers by Andrew Longbottom and Sean Holly (1985a) and by Alvaro Escribano (1985), which they regard

⁹The standard error of estimate for our regression is 5.54 percent; for HE's regression 9, it is 10.86 percent; for that in their table 3, it is 15.5 percent (it is given as 1.55, but that has to be multiplied by 10 to be comparable to the other two, since the long-run dependent variable is $0.1m$).

¹⁰HE report the standard error of estimate as 1.424 percent. Our attempted duplicate has a standard error of estimate of 1.50 percent.

as producing "significant improvements on our [i.e., HE's] 1983 model" (HE, p. 24) but still leaving room for further improvement, as in HE's revised models. In a slightly later paper by Longbottom and Holly (1985b), which HE do not refer to but which, like the earlier paper, uses HE's proposed procedures, the final summary is: "In this paper we have re-examined Friedman and Schwartz's work on UK monetary trends in the light of the methodological criticisms of Hendry and Ericsson. In contrast to Hendry and Ericsson we are able to find empirical support for the claims that Friedman and Schwartz make about the form and long run stability of the demand for money function in the UK since 1878" (Longbottom and Holly, 1985b p. 19).

We hasten to add that we regard the Longbottom and Holly papers no less an example of formal econometric analysis carried to extremes than HE's papers. Their confirmation of our results does not increase our confidence in our results any more than HE's assertion that we have failed "to present statistical evidence pertinent to their [our] main claims about the United Kingdom" (p. 32) weakens our confidence in them. Our confidence in our results derives, as we have stressed repeatedly, from a much broader base and would not be justified if their analysis was all we had to rely on.

F. *The Real Proof of the Pudding*

Before ending this section, in which we have expressed so much skepticism about HE's approach, we should indicate what evidence would persuade us that we are wrong and they are right. The answer is straightforward. A persuasive test of their results must be based on data not used in the derivation of their equations. That might mean using their equations to predict the same kind of phenomena for other countries, or for a future or earlier period for the United Kingdom, or deriving testable implications from their equations for other variables, such as exchange rates, term structure of interest rates, or still other phenomena we are not imaginative enough to list. Similarly, that is the only kind of evi-

dence that we would regard as persuasive with respect to the validity of our own results.¹¹

V. Conclusion

Hendry and Ericsson describe their paper as an evaluation of "an empirical model of U.K. money demand developed by Friedman and Schwartz in *Monetary Trends*" (HE, p. 8). Viewed from that point of view, seldom can a mountain have labored so hard and produced so small a mouse. After years of experiments, HE's econometric techniques produced a series of models that confirm some of our principal results, contradict none, and are less successful than our equations in terms of their own criterion of variance-dominance.

But their paper is mislabeled. It is not in any relevant sense an evaluation of our "empirical model of U.K. money demand." They use one out of our hundreds of regressions as a peg on which to hang an exposition of a set of sophisticated econometric techniques designed for a purpose and embodying a methodological approach very different from ours. Their regressions are designed to explain the short-term adjustment process as well as the long-term relation. We had no such ambitious aim.

We are incompetent to judge the adequacy of their techniques for estimating the short-term adjustment process. However, we do not regard their statistical tests as demonstrating the validity of their statistical estimates. Their estimates are the end result of trying a large number of alternative hypotheses on a single body of data. As a result, it is impossible to specify how many

¹¹Our initial draft of *Monetary Trends* was based on U.S. data only. We expanded it to include the United Kingdom precisely in order to test our initial generalizations with data other than those used in deriving them. This is of course a never-ending process; the generalizations in the published book in their turn should be tested against data not used in deriving them. One minor such test extending the period originally covered is that in Friedman (1988 footnote 16). A regression for the United States for 1886-1985 using annual data gives essentially the same results as our phase-average equation for 1873-1975.

"degrees of freedom" have been used up in the process of reaching the final equations presented, or, put differently, to estimate the probabilities that their results could have arisen from chance. For that, one needs, in their words, "the underlying (and unknown) error variances," not "the *estimated residual variance*" on which they rely (HE, footnote 14). As already indicated, the real proof of their pudding is whether it produces a satisfactory explanation of data not used in baking it—data for subsequent or earlier years, for other countries, or for other variables. One example of such a test, in a physical-science context, is given in the Appendix. That example dramatically illustrates how misleading a multiple regression can be for predictive purposes, even though it satisfies all the standard tests.

APPENDIX: A CAUTIONARY TALE ABOUT MULTIPLE REGRESSIONS

*(This addendum was written by
Milton Friedman)*

My skepticism about relying on a single multiple regression that results from the HE approach traces back to an experience I had in 1944 or 1945 when I was engaged in war research as a member of the staff of the Statistical Research Group of Columbia University.

One of my assignments was to serve as a statistical consultant to a number of projects seeking to develop an improved alloy for use in airplane turbo-superchargers and as a lining for jet engines. The goal was to develop alloys that could withstand the highest possible temperature, since the efficiency of a turbine (or its equivalent) rises very rapidly with the temperature at which it can safely operate. I served as something of a clearing agency for the results of the various experiments in progress, as an adviser on statistical design of experiments, and as an analyst of the results, producing a fairly regular newsletter on these matters for the experimenters.

The procedure in testing an experimental alloy was to hang a specified weight on a standard turbine blade made from the alloy,

put it in a furnace capable of generating a very high temperature, and measure the time it took for the blade to break. At one point, I combined the test data from all the separate experiments and engaged in precisely the kind of analysis that HE recommended. I ended up with a single proposed regression that expressed time to fracture as a function of stress, temperature, and variables describing the composition of the alloy. I assured myself that the equation was consistent with metallurgical theory.

The major problem then, trivial now, was to compute the parameters of the equation and the associated test statistics. That was the age of the desk electric—not electronic—calculators and the Dolittle method of computing regressions. The labor involved in that method increases exponentially with the number of independent variables. For the number I wanted to use, we estimated that it would take three months for one of our highly skilled operators to calculate the equation. Fortunately, we discovered that there was one large-scale computer in the country that could perform our calculations: the experimental Mark I (or something like that) at Harvard, itself not electronic but built from a large number of IBM card-sorting machines housed in an enormous air-conditioned gymnasium. We were granted time on the machine to perform our calculations. Today's statisticians will be interested to know that, not counting data insertion, it took 40 hours to calculate a regression that I can now calculate on my desktop computer in less than 30 seconds—my favorite story to illustrate what has happened to our computer power.

I was delighted with the calculated regression. It had a high multiple correlation, low standard error of estimate, and high t values for all of the coefficients, and it satisfied every other test statistic that I knew of more than 40 years ago. I immediately set to work to create some new and better alloys. In constructing such alloys, I had to go outside the joint range of my sample set of independent variables, but I was careful to stay as close as I could and to be within the limits used in prior experiments for each variable separately. The technical details are

irrelevant for the present purpose, and I could no longer reproduce them in any event.

The bottom line is that I ended up constructing two new alloys (which with hope combined with caution, I named F-1 and F-2). According to the calculated regression, each would take several hundred hours to rupture at the very high temperature I proposed to test them at, a sizable multiple of the best recorded time for any previous alloy. This was physics, not economics, so I did not have to wait years to see whether the predictions from my equation were correct. I phoned an MIT lab that was working on alloys of a similar type and asked them to cook up and test my two alloys. I was sufficiently skeptical—or perhaps just cautious—so that I was careful not to tell them what to expect. A few days later they phoned the results: my two alloys had ruptured in something like 1–4 hours, a much poorer outcome than for many prior alloys. F-1 and F-2 were never heard of again.

Ever since, I have been extremely skeptical of relying on projections from a multiple regression, however well it performs on the body of data from which it is derived; and the more complex the regression, the more skeptical I am. In the course of decades, that skepticism has been justified time and again. In my view, regression analysis is a good tool for deriving hypotheses. But any hypothesis must be tested with data or non-quantitative evidence other than that used in deriving the regression or available when the regression was derived. Low standard errors of estimate, high t values, and the like are often tributes to the ingenuity and tenacity of the statistician rather than reliable evidence of the ability of the regression to predict data not used in constructing it.

REFERENCES

- Escribano, Alvaro, "Non-linear Error-Correction: The Case of Money Demand in the U.K. (1878–1970)," mimeo, University of California, San Diego, 1985.
- Friedman, Milton, *A Theory of the Consumption Function*, Princeton: Princeton University Press, 1957.
- _____, "Money and the Stock Market," *Journal of Political Economy*, April 1988, 96, 221–45.
- _____, and Kuznets, Simon, *Income from Independent Professional Practice*, New York: National Bureau of Economic Research, 1945.
- _____, and Schwartz, Anna, J., *A Monetary History of the United States, 1867–1960*, Princeton: Princeton University Press (for the National Bureau of Economic Research), 1963.
- _____, and _____, *Monetary Trends in the United States and the United Kingdom: Their Relation to Income, Prices, and Interest Rates, 1867–1975*, Chicago: University of Chicago Press (for the National Bureau of Economic Research), 1982.
- Hendry, David F. and Ericsson, Neil R., "An Econometric Analysis of U.K. Money Demand in *Monetary Trends in the United States and the United Kingdom* by Milton Friedman and Anna J. Schwartz," *American Economic Review*, March 1991, 81, 8–38.
- Hotelling, Harold, "Review of *The Triumph of Mediocrity in Business* by Horace Secrist," *Journal of the American Statistical Association*, December 1933, 28, 463–5.
- _____, Letter to the Editor, *Journal of the American Statistical Association*, June 1934, 29, 198–9.
- Klein, Benjamin, "The Payment of Interest on Commercial Bank Deposits and the Price of Money: A Study of the Demand for Money," unpublished Ph.D. Dissertation, University of Chicago, 1970.
- Longbottom, Andrew and Holly, Sean, (1985a) "Econometric Methodology and Monetarism: Professor Friedman and Professor Hendry on the Demand for Money," London Business School Discussion Paper No. 131, February 1985.
- _____, and _____, (1985b) "Monetary Trends in the UK: A Reappraisal of the Demand for Money," London Business School Discussion Paper No. 147, April 1985.
- Secrist, Horace, Letters to the Editor, *Journal of the American Statistical Association*, June 1934, 29, 196–8, 200.

The Failure of Competition in the Credit Card Market

By LAWRENCE M. AUSUBEL*

The bank credit card market, containing 4,000 firms and lacking regulatory barriers, casually appears to be a hospitable environment for the model of perfect competition. Nevertheless, this article reports that credit card interest rates have been exceptionally sticky relative to the cost of funds. Moreover, major credit card issuers have persistently earned from three to five times the ordinary rate of return in banking during the period 1983–1988. The failure of the competitive model appears to be partly attributable to consumers making credit card choices without taking account of the very high probability that they will pay interest on their outstanding balances. (JEL 315, 612).

This article presents and discusses a collection of data which is paradoxical within the paradigm of perfect competition. The market studied, the bank credit card industry in the United States, contains literally 4,000 firms who sell a relatively homogeneous good to 75 million consumers. The ten largest firms account for only about two-fifths of market share. Firms have historically operated without regulatory barriers to conducting business across state lines—and at least 20 firms aggressively solicit business on a national scale. Firms have also operated in the virtual absence of price regulations for most of a decade. There do not appear to be any particularly con-

strained inputs, significant sunk costs, or significant barriers to entry. Finally, there is no evidence of any explicit collusion on price or quantity.

Given such a favorable market description, or one not even half so optimistic, many economists would prefer to presume that the market must behave as a competitive spot market in continuous equilibrium. It is the purpose of this article to argue that this presumption is empirically unjustified in the market for bank credit cards in the 1980's. Section I outlines the market structure of the bank credit card industry. Section II offers empirical evidence of extreme price stickiness in credit card interest rates. Section III provides direct profit data on the industry, arguing that the 50 largest credit card issuers have earned from three to five times the ordinary rate of return for the banking industry during the period 1983–1988. Section IV examines profits over a larger sample of banks and a longer time period. Section V presents additional data on resales of credit card portfolios between banks, suggesting that the extraordinary profits exist *ex ante* as well as *ex post* (and that bankers expect the profitability to persist). Section VI explores some theoretical explanations for price stickiness and supra-normal profits. Section VII calculates what would be "competitive" interest rates. Section VIII briefly discusses the extent of welfare loss in the market and the merits of regulation to correct market failure. Conclusions are presented in Section IX.

*Department of Managerial Economics and Decision Sciences, J. L. Kellogg Graduate School of Management, Northwestern University, Evanston, IL 60208. The author acknowledges the support of the Kellogg School's Banking Research Center, The Lynde and Harry Bradley Foundation, and the C. V. Starr Center at New York University and appreciates the diligent research work of Gail Eynon and Paul Palmer. I thank Alan Blinder, Charles Calomiris, Raymond Deneckere, Peter Diamond, Stuart Greenbaum, Robert Johnson, Charles Kahn, Robert Porter, and three anonymous referees for helpful comments. I also thank seminar participants at the American Economic Association Meetings, the Econometric Society Meetings, the NBER Economic Fluctuations Conference, the Northwestern University Summer Industrial Organization Conference, the Federal Reserve Bank of Chicago, New York University, Princeton University, and the University of Delaware. Special thanks are also due to the officers of 21 major banks who cooperatively responded to my requests for data.

I. The Bank Credit Card Market: Is 4,000 Enough for Competition?

Credit cards are the currency of late 20th-century America. The aggregate charge volume on plastic in the United States was estimated at \$375 billion in 1987.¹ Almost half of this total—\$165 billion in volume—was charged on MasterCard and Visa credit cards (the primary focus of this article), and volume was growing at well over 10 percent per year.² The remaining volume arose largely from similar credit cards (e.g., the Discover and Optima cards), “travel and entertainment” cards (e.g., the American Express card), and retail cards (e.g., department store and oil company cards).

Borrowing via credit cards (and all consumer borrowing) is also significant and has been even more of a growth industry. Outstanding U.S. balances on revolving credit accounts equaled \$203 billion at year-end 1989, up from only \$70 billion in 1982.³ More than \$130 billion of this total consisted of MasterCard and Visa balances, more than a threefold increase from 1982, and bank card balances were still increasing at more than a 15-percent annual rate.⁴ Overall outstanding consumer installment credit balances in the United States reached \$717 billion, up from \$356 billion in 1982;⁵ it is worth observing that many of the considerations explicitly discussed here in con-

nection with the credit card industry apply also to other forms of consumer borrowing (especially other unsecured credit).

If Visa and MasterCard were the relevant levels of business to examine, then two firms would control a substantial part of the credit card market. However, most relevant business decisions are made at the level of the issuing bank. Individual banks own their cardholders' accounts and determine the interest rate, annual fee, grace period, credit limit, and other terms of the accounts. (Only charges such as the “interchange fee” from the merchant's bank to the cardholder's bank are standardized, and the cardholder's bank appears only to break even on such charges. Moreover, there is absolutely no indication that the MasterCard and Visa organizations serve to facilitate collusion on other prices.⁶) In essence, MasterCard International and Visa U.S.A. are organizations largely irrelevant to this discussion; “firms” will henceforth refer to the issuing banks.

The market for MasterCard and Visa cards, thus, is relatively unconcentrated. The top ten firms control only about two-fifths of the market, and the next ten firms control only one-tenth of the market (see Table 1). Moreover, the market is exceptionally broad. A bank that ranked number 100 in 1987 still had approximately 160,000 active accounts, \$125 million in outstanding balances, and \$250 million in annual charge volume.⁷

Unlike most aspects of American banking, the credit card business has historically operated free of interstate banking and

¹Moreover, Americans were estimated to have made 9.1 billion credit card transactions in 1987 (*The Nilson Report*, Number 428, May 1988, p. 5).

²U.S. volume in 1987 consisted of \$138 billion in sales slips (i.e., charged goods and services) and \$27 billion in cash advances. Visa accounted for 59 percent of this value and MasterCard accounted for the remaining 41 percent. (*The Nilson Report*, Number 422, February 1988, p. 6, and Number 423, March 1988, pp. 4–5).

³Federal Reserve Board's series of Consumer Installment Credit, as published in *Federal Reserve Bulletin*, April 1990, table 1.55, line 15 (and previous issues).

⁴*Federal Reserve Bulletin*, April 1990, table 1.55, lines 16, 19, and 21. Revolving credit held by commercial banks, savings institutions, and pools of securitized assets consists almost entirely of MasterCard and Visa balances.

⁵*Federal Reserve Bulletin*, April 1990, table 1.55, line 1 (and previous issues).

⁶Moreover, the observed interest rate behavior does not seem to fit the conventional view of collusive pricing. Around 1985, three major issuers (Chase Manhattan, Manufacturers Hanover, and Maryland Bank) reduced their interest rates on standard cards to the 17.5–17.9-percent range. Far from this triggering an industry price war, other major issuers (e.g., Citibank and First Chicago) steadfastly maintained 19.8-percent rates on most accounts, without apparent detriment to their customer bases. Finally, in the spring of 1989, the three price-cutters announced rate increases, apparently finding *without facing retaliation* that the earlier cuts had been unprofitable (*The New York Times*, April 27, 1989, p. 32; *Wall Street Journal*, March 22, 1989, p. B1).

⁷*The Nilson Report*, Number 406 (June 1987), p. 7.

TABLE 1—TOP TEN ISSUERS OF MASTERCARD AND VISA CARDS, 1987

Bank	Number of accounts	Percentage market share (by number of accounts)	Outstanding balances (\$ billion)	Percentage market share (by outstanding balances)
Citibank	10,000,000	8.4	\$15.3B	16.3
Chase Manhattan Bank	5,000,000	4.2	\$5.4B	5.8
Bank of America	4,800,000	4.0	\$5.2B	5.5
First Chicago	4,500,000	3.8	\$4.6B	4.9
Manufacturers Hanover	3,300,000	2.8	\$2.0B	2.1
Wells Fargo Bank	1,800,000	1.5	\$2.8B	3.0
Maryland Bank	1,800,000	1.5	\$1.7B	1.8
Marine Midland Bank	1,700,000	1.4	\$1.4B	1.5
Chemical Bank	1,500,000	1.3	\$1.3B	1.4
Associates National Bank	1,200,000	1.0	\$1.0B	1.1
Top ten	35,600,000	30.0	\$40.7B	43.4
Second ten	11,500,000	9.7	\$9.0B	9.6
Total	118,900,000	100.0	\$93.9B	100.0

Sources: Individual banks' numbers of accounts surveyed by *American Banker*, (March 1, 1988, pp. 1–2) and *Credit Card News* (August 15, 1988, pp. 4–16); total number of accounts from *Nilson Report* (Number 406 [June 1987], p. 4). Individual banks' outstanding balances based on *American Banker* (September 21, 1987, p. 43 [call report data]) and *Nilson Report* (Number 406 [June 1987], pp. 4–5); total outstanding balances from *Federal Reserve Bulletin* (December 1988, table 1.55; revolving credit outstanding at commercial banks and savings institutions, minus loans outstanding at Greenwood Trust Co. [Discover card] and American Express Centurion Bank). Data reported for December 31, 1986, adjusted for acquisitions effective in 1987. Conflicts between sources were resolved using best available information.

branch banking restrictions. Indeed, the largest issuers today conduct truly national businesses. For example, Maryland Bank (ranked number seven in Table 1) conducts business in all 50 states and has only five percent of its accounts in its home state.⁸ The only states where more than five percent of its business is concentrated are California (10.7 percent), Texas (6.7 percent), Pennsylvania (6.0 percent), and New Jersey (5.8 percent).

In the past, credit card issuers were constrained by state usury laws. However, the U.S. Supreme Court's December 1978 *Marquette* decision paved the way for the practical elimination of price regulations.⁹ The

Court held that only the usury ceiling of the state in which the bank is located, and not that of the state in which the consumer is located, restricts the interest rate the bank may charge. This gave banks the option of shifting their credit card operations to wholly owned subsidiaries situated in states without usury laws. By 1982, amid *Marquette*-created bank pressure and historically high market interest rates, most leading banking states had relaxed or repealed their interest rate ceilings. Meanwhile, South Dakota and Delaware had established themselves as attractive homes-away-from-home for credit card issuers. While a number of states maintain binding usury laws at this writing (most notably, Arkansas, with a ceiling of five percentage points above the Federal Reserve discount rate), essentially all major issuers can pursue business in those states free of restric-

⁸"Prospectus for Maryland Bank, N.A., Credit Card Trust 1987-A," December 9, 1987, pp. 17–18.

⁹*Marquette National Bank v. First of Omaha Service Corporation*, 439 U.S. 299 (1978). The *Marquette* decision applies to credit cards issued by nationally chartered banks, but not to retail cards (e.g., oil company credit cards). The decision explicitly permits banks to "export" their interest rates; banks have interpreted this also to permit the "export" of annual fees and

other customer fees. At this writing, at the behest of the Iowa Attorney General, courts are considering whether this rule does indeed apply to fees.

tion. It is fair to say that the bank credit card market in the United States was functionally deregulated in 1982.

II. Credit Card Interest Rate Behavior

A. Sticky Interest Rates

The cost of funds is obviously the primary determinant of the marginal cost of lending via credit cards, and it is usually the only component of marginal cost that varies widely from quarter to quarter. Thus, a model of continuous spot market equilibrium would predict a substantial degree of connection between the interest rate charged on credit cards and the banks' cost of funds. However, Figure 1, which compares credit card interest rates with the cost of funds, displays stark empirical rejection of this prediction. Credit card interest rates were highly sticky during the period 1982–1989 and, in fact, were virtually constant.¹⁰

In this section, credit card interest rates are captured by two distinct sets of data: one aggregated and one disaggregated. The first set of data is the *Federal Reserve Bulletin* series for credit card interest rates, based on the Federal Reserve Board's quarterly survey of banks. Reported are arithmetic averages of each bank's "most common" rate charged during the first week of each mid-quarter month.¹¹ This series is plotted in Figure 1. The second set of data (and much of the empirical discussion of this and the next section) is derived from the author's own bank credit card survey (BCCS) of 58 of the largest bank issuers of credit cards. The first mailing (21 responses) asked primarily for pricing and cost data; it generated a quarterly interest rate series for 17 credit card issuers and an annual loan-loss series for 10 issuers. The follow-up

mailing (11 responses) included a request for direct profit calculations, which were provided by seven banks. Appendix A provides details of the construction of the BCCS. Table 2 includes the size distribution of banks that reported data. Respondents were promised anonymity.

The most aesthetically pleasing way for an economist to determine the cost of funds is to "let the market decide it." In the case of credit cards, this is feasible because of the phenomenon of credit card securitization. Consistently, during 1987–1989, credit-card-backed securities offered yields in the vicinity of 0.75 percent above those of Treasury securities with comparable maturities.¹² Meanwhile, the Visa systemwide average cardholder payment rate (i.e., cardholder payments as a percentage of outstanding balances) ranged from 13 to 17 percent per month during the years 1983–1987, implying an average maturity for credit card receivables of 6–8 months.¹³ To be conservative, I will define the cost of funds to equal the one-year Treasury bill yield¹⁴ plus 0.75 percent, averaged over each quarter. This series is also plotted in Figure 1.

The proposition that interest rates are sticky can be formally supported by regressing credit card interest rates on the cost of

¹²See, for example, "Credit Card Bonds are Hot, but Maybe Stingy on Yield," *Wall Street Journal*, April 16, 1990, p. C1; *Credit Card News*, Volume 1, Number 3 (June 15, 1988), p. 2, and Volume 1, Number 14 (November 15, 1988), p. 7; *Credit Card Management*, May/June 1988, p. 34.

¹³The source of the systemwide cardholder payment rate is Standard & Poor's *Asset-Backed Securitization Credit Review*, March 16, 1987, p. 19. Individual banks' prospectuses have reported cardholder payment rates of 9–23 percent per month, never implying an average maturity of more than one year (see list of prospectuses in Appendix B). This impression was substantiated by a trade-publication report quoting the chairman of FCC National Bank (First Chicago's Delaware credit card subsidiary, listed fourth in Table 1) as saying that his bank finances its credit card portfolio with a variety of financial instruments with combined maturities equivalent to a 145-day duration (*Credit Card News*, March 15, 1989, p. 2).

¹⁴*Federal Reserve Bulletin*, April 1990, table 1.35, line 21 (and previous issues).

¹⁰Indeed, the highest value reported in the *Federal Reserve Bulletin* series in the period 1982–1989 is 18.85 percent (first quarter, 1985) and the lowest value reported is 17.77 percent (fourth quarter, 1988).

¹¹Federal Reserve Board's G.19 statistical release, April 5, 1990; *Federal Reserve Bulletin*, April 1990, table 1.56, line 4 (and previous issues).

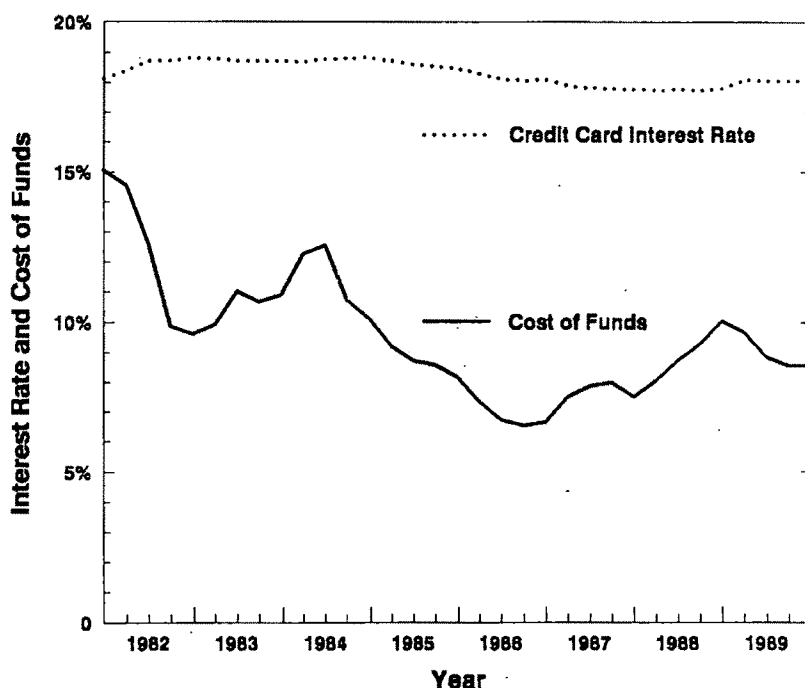


FIGURE 1. STICKY CREDIT CARD INTEREST RATES, 1982-1989

Notes: Credit card interest rate is the quarterly Federal Reserve System series; cost of funds is the quarterly one-year Treasury bill yield plus 0.75 percent.

funds. First, using the Federal Reserve series, an aggregate credit card interest rate is regressed on its own lagged value, the lagged cost of funds, and a constant. Second, a more thorough regression can be run using the author's BCCS series for the 17 individual banks: each bank's credit card interest

rate is regressed on its own lagged value, the lagged cost of funds, and a dummy variable for that bank. The results of these linear regressions are reported in Table 3. Note that, in the second regression, every coefficient has a *t* statistic of at least 6, while inclusion of additional variables with

TABLE 2—SIZE OF BANK CREDIT CARD ISSUERS FOR WHICH DATA ARE REPORTED

1987 ranking by number of active accounts (1 = largest)	Number of banks with BCCS reports enabling computation of profits	Number of banks with call reports enabling computation of profits	Number of banks with prospectuses enabling estimation of profits	Number of banks with BCCS reports of interest rate series
1-10	1	3	4	2
11-20	2	3	0	5
21-30	2	1	1	3
31-40	1	1	2	1
41-50	1	1	1	4
51+	0	0	0	2
Total:	7	9	8	17

Sources: Author's bank credit card survey (BCCS), consolidated reports of condition and income (call reports), and prospectuses and registration statements. Ranks according to *The Nilson Report*, Number 406 (June 1987), pp. 6-7.

TABLE 3—ORDINARY LEAST-SQUARES REGRESSION
OF CREDIT CARD INTEREST RATE ON COST OF
FUNDS AND LAGGED CREDIT CARD INTEREST RATE
(QUARTERLY, 1982–1987)

Variable	Federal Reserve Board survey data	Bank credit card survey data
COST OF FUNDS ₋₁	0.0422 (0.00584)	0.0540 (0.00896)
CREDIT CARD INTEREST RATE ₋₁	0.895 (0.0444)	0.685 (0.0326)
Constant	1.51 (0.807)	
Bank-1 dummy		5.75 (0.659)
Bank-2 dummy		5.11 (0.594)
Bank-3 dummy		6.38 (0.719)
Bank-4 dummy		5.79 (0.657)
Bank-5 dummy		5.70 (0.651)
Bank-6 dummy		4.18 (0.500)
Bank-7 dummy		5.69 (0.645)
Bank-8 dummy		5.12 (0.595)
Bank-9 dummy		5.61 (0.644)
Bank-10 dummy		5.44 (0.634)
Bank-11 dummy		5.00 (0.589)
Bank-12 dummy		5.12 (0.595)
Bank-13 dummy		4.99 (0.586)
Bank-14 dummy		4.88 (0.577)
Bank-15 dummy		5.50 (0.636)
Bank-16 dummy		5.22 (0.593)
Bank-17 dummy		6.17 (0.707)
Number of observations:	24	408
R ² :	0.96	0.937
Durbin h:	-0.69	0.10

Notes: CREDIT CARD INTEREST RATE is the dependent variable in each regression. COST OF FUNDS is defined as the yield on one-year Treasury bills plus 0.75 percent. Observe that there is no cross-firm variation in this variable, so that year dummy variables cannot be included in the second regression equation. Banks included in the author's bank credit card survey were assured anonymity. Numbers in parentheses are standard errors.

other lags tended to cause some coefficients to become insignificant. To aid in comparing the results of the two regressions, the Fed series is used only for the period 1982–1987; using 1982–1989 data yields the same conclusions.¹⁵

The coefficient on the cost of funds, while statistically significant in each of the two regressions, is economically insignificant. Whereas a competitive-spot-market model would predict a coefficient near 1, the regressions using aggregated and disaggregated data yielded coefficients of only 0.042 and 0.054, respectively. It takes many years for the price to adjust to changes in marginal cost when the rate of adjustment is only on the order of 5 percent per quarter.

B. Nonprice Competition .

The credit card industry has defended its high interest rates in the mid-to-late 1980's, in part, by asserting that the increased spread between the credit card interest rate and the cost of funds had been caused by an increase in the industry's rate of bad loans. The loan-loss data from the author's BCCS indicate that, in the period 1982–1987, the charge-off rate actually did increase roughly coincident with the increase in the interest rate spread (see Table 4). However, higher loan losses are an explanation for the higher interest rate spreads only if one believes that the latter are solely determined by costs. If credit card interest rates are determined otherwise, then *the causation may run in the reverse direction*: an increased interest rate spread may cause an increase in charge-offs.

Suppose, for example, that a bank can select both its interest rate and the default risk of its marginal customer. By choosing a higher marginal default rate, the bank increases its total number of loans but also its charge-off rate (the average default rate). Suppose that the bank first selects its interest rate and then its marginal default rate. Profit maximization requires the bank to set

¹⁵With the 1982–1989 *Federal Reserve Bulletin* series, one obtains a coefficient of 0.0439 on COST OF FUNDS₋₁, a coefficient of 0.864 on CREDIT CARD INTEREST RATE₋₁, and a constant of 2.06.

TABLE 4.—LOAN LOSSES ON CREDIT CARDS OF
TEN SURVEY BANKS DURING 1976–1987
(FEWER THAN NINE DURING 1976–1978 AND 1986–1987)
AND THE SPREAD BETWEEN CREDIT CARD
INTEREST RATES AND THE COST OF FUNDS

Year	Average charge-off rate (percentage)	Interest rate spread (percentage)
1976	1.15	10.57
1977	0.99	10.36
1978	1.27	8.11
1979	1.44	5.78
1980	2.04	5.08
1981	1.48	2.94
1982	1.67	6.42
1983	1.32	8.79
1984	1.36	7.69
1985	1.94	9.82
1986	3.01	11.55
1987	2.60	10.43

Source: Author's bank credit card survey (Appendix A, Table A1, questions 7 and 1).

its marginal default rate equal to the difference between the interest rate it charges and the marginal cost (net of defaults) of lending funds. (The net marginal cost should equal the cost of funds plus a constant that is fairly stable in the short run.) Thus, the prediction is that an optimizing bank should set its marginal default rate equal to the interest rate spread plus a constant.

Suppose now that there is an independent reason why credit card interest rates fail to fall when general market interest rates decline (for example, see Section VI, below). The logic of the previous paragraph dictates that loan losses will subsequently increase. If firms do not compete and drive price down toward marginal cost, they are likely instead to compete and drive marginal cost up toward price,¹⁶ in the form of issuing cards to less credit-worthy customers.

¹⁶A related argument was made in the context of airline regulation. George W. Douglas and James C. Miller (1974) argued that the Civil Aeronautics Board's price regulations, at a time when the introduction of jet engines reduced the fundamental cost of air transportation, led airlines to compete and drive their costs up to price by placing fewer passengers on a given airplane. The arguments differ in two important ways. First, in the airline industry, price rigidity may have been caused by price regulation, whereas with credit

III. The *Ex Post* Profitability of the Credit Card Market

As seen in Section I, the credit card market of the 1980's possessed most of the usual prerequisites for invoking the model of perfect competition. A perfectly competitive model would at least predict zero long-run economic profits for "marginal" firms. Moreover, since free entry into the industry is possible and no input appears to be in scarce supply,¹⁷ there is no credible source of rents to distinguish "inframarginal" firms from "marginal" firms. Thus, the competitive model would predict that all credit card issuers earn zero long-run economic profits. Many models of imperfect competition which preserve the free-entry assumption would also yield the zero-profit prediction.

By way of contrast, the interest rate stickiness documented in the previous section suggests that credit cards must become extraordinarily profitable whenever the cost of funds drops. Indeed, in this section, I will present a rather paradoxical set of data which indicates that returns from the credit card business were several times greater than the ordinary rate of return in banking during the years 1983–1988.

At the same time, this profitability data will help to assure that the above evidence of interest rate stickiness has been correctly interpreted. One might have thought to argue that price rigidity is consistent with

cards, there is price stickiness despite a deregulated environment. Second, under regulation, the airlines apparently competed away their profits.

¹⁷Free entry is a reasonable depiction of a credit card market in which 4,087 banks (and other deposit institutions) already issued their own Visa cards and a similar (largely overlapping) number issued their own MasterCard in September 1987. All of these institutions could legally offer accounts to customers anywhere in the United States. Nonmember institutions could join the Visa system by paying a fairly trivial entry fee: six dollars per million dollars in assets, plus one thousand dollars, according to a Visa official. (Only the assets of the subsidiary that issues the cards, and not the assets of the holding company, are figured into this formula.) Furthermore, it would seem strained to argue either that adjustment to the "long run" requires many years or that some input is in scarce supply, given the deluge of credit card solicitations made by banks in recent years.

competitive spot markets, if unobservable increases in quality exactly offset reductions in factor costs. The profitability data enable one to dismiss this possibility: profits, in fact, dramatically rose at the time that the cost of funds dropped.

It is possible to object to the following analysis on several grounds. First, the data reported, by their very nature, represent *ex post* profits. Perhaps (especially since the sample period is during a cyclical boom) the observed profits are merely a very favorable realization of a random variable whose *ex ante* returns were quite ordinary. Second, it might be thought that, while the credit card market was extremely profitable in the years 1983–1988, the market has now equilibrated and henceforth normal returns will be observed. Third, the profitability figures might be derived from accounting data that either are being misinterpreted or are systematically misstating true economic profits.

I consider each of these concerns elsewhere in the paper. In Section IV, I briefly discuss an additional source of evidence (the Federal Reserve System's functional cost analysis), which, while significantly less reliable than the other data (in this author's opinion), gives profits over a longer period that includes the previous cyclical downturn. In Section V, I introduce another independent set of data which examines resale prices of credit card portfolios between banks and finds that they trade at large premia. The latter data indicate that *ex ante* returns from credit cards are quite large and, since they are based on market valuations, should help allay any fears that the accounting data are being misinterpreted. Finally, it should be recalled from Table 4 that the interest rate spread was quite healthy except for a brief period around 1981 and that this brief spell of unprofitability can be attributed to banks not having yet established credit card subsidiaries exploiting the Supreme Court's *Marquette* decision. This episode does not seem likely to be repeated.

The *ex post* profit data reported and discussed in this section originate from three independent sources and were assembled by the author.

Bank credit card survey: The author's follow-up survey yielded profit calculations performed directly by executives of seven of the 50 largest bank issuers of credit cards.

Call reports: Profitability data for another nine of these issuers were extracted from call reports filed by the banks with the FDIC.

Prospectuses: Partial data on profitability for an additional eight large banks were obtained from filings with the SEC in connection with the sale of credit-card-backed securities.

Respondents to the author's survey were promised anonymity (but details of the construction are provided in Appendix A). The call reports and prospectuses are part of the public record. Table 2 reports the size distribution of banks included in each of the survey, call report, and prospectus samples.

A. An Illustrative Profit Calculation

As will be detailed in the next two subsections, earnings in the banking industry are usefully expressed as a percentage of assets: returns on assets are linked with returns on equity by the banking system's capital requirements. Before reporting summary profit figures for 15 and estimates for eight of the 50 largest issuers, I will examine in detail the components of revenues and costs for one individual credit card issuer. I consider here Maryland Bank, N.A. (MBNA), the Delaware-based credit card arm of MNC Financial, which is ranked seventh in Table 1.¹⁸ This institution was selected because more public information exists on its credit card operations than on any other bank's: MBNA, which is required to file its own call report, has credit card loans exceeding 92 percent of its assets, and it has also made several credit-card-backed securities offerings.

¹⁸MNC Financial is the 39th largest U.S. bank holding company and the corporate parent of Maryland National Bank, the largest commercial bank in Maryland. MBNA was founded in Newark, Delaware, in 1982, apparently to avoid Maryland's usury law. See also the text near footnote 9.

TABLE 5—COMPONENTS OF PROFITS FOR MARYLAND BANK, N.A.

Components	1985	1986	1987
Finance charges	16.66%	14.92%	13.21%
Annual fees	1.40%	1.58%	1.29%
Other customer charges	1.10%	1.42%	1.17%
Interchange fees	3.06%	3.00%	2.92%
Total revenue:	22.22%	20.92%	18.60%
Interest expenses	9.57%	7.80%	7.13%
Noninterest expenses	4.47%	4.71%	4.87%
Net charge-offs	1.09%	1.77%	1.80%
Total cost:	15.13%	14.28%	13.80%
Return on assets (pretax profits expressed as a percentage of outstanding balances)	7.09%	6.63%	4.80%

Sources: Consolidated reports of condition and income (call reports), prospectuses, and registration statements for Maryland Bank, N.A.

MBNA's credit card operations (and their profitability) are fairly typical of major issuers, with the exception that the bank has stressed the concept of "affinity credit cards," whereby cards are marketed to members of professional organizations, fraternal orders, and cause-related groups (with the organizations' endorsements). As a consequence, its interest rates are somewhat lower and its customers are somewhat more credit-worthy than average. Indeed, it may interest readers that, during the period when this article was undergoing the journal's review process, MBNA's marketing agent proposed to establish an official American Economic Association Visa card. This card would have carried a \$20 annual fee (\$40 for a gold card) and an 18.9-percent annual fee; the AEA would have received \$1 for each account opened, \$3 for each account renewed, and \$0.25 per retail transaction. MBNA's agent estimated that 1,000 cards would be issued, generating \$13 in revenue per card per year for the AEA. However, the AEA's executive committee, concerned that the AEA "would be viewed as endorsing a specific credit card by entering into such a contract," voted against establishing the affinity card program.¹⁹

¹⁹Draft minutes of the March 23, 1990, meeting of the AEA Executive Committee; Report of the Secre-

An item-by-item profit calculation for MBNA is displayed in Table 5. As is typical for credit card issuers, the single largest component of revenue is the finance charge (which, for MBNA, derives from annual percentage rates of 14.5–18.9 percent, depending on the account). Despite the fact that the bank provides a 25-day grace period during which no finance charge is assessed if the account balance is paid in full, more than 80 percent of the bank's credit card outstanding balances do accrue interest.²⁰ The drop in finance-charge revenues displayed in Table 5 is largely attributable to the bank's decision to reduce the interest rates on some of its accounts during 1985–1987.

MBNA also derives direct customer revenues from the annual fee and other customer charges (e.g., \$15 late payment, over-limit, and returned-check charges). Indirect revenues are derived from the interchange fee, the portion of the merchant discount that is paid to the customer's bank. It is worth reemphasizing that the price sched-

tary to the Executive Committee. I thank Orley Ashenfelter and C. Elton Hinshaw for providing this information.

²⁰A good rule of thumb mentioned in credit card trade publications is that 90 percent of an issuer's overall outstanding balances accrue interest. See the discussion in Section VI-C.

ules that determine direct customer revenues are set entirely at the bank level; only the interchange fee is set systemwide by the MasterCard and Visa organizations.

Costs can be divided among interest expenses, operating expenses, and loan losses. Interest expense is determined by market interest rates and thus is relatively uniform across banks, as a percentage of outstanding balances. Noninterest expenses, which include employee salaries, occupancy, equipment, and data processing, are also available from the banks' direct reports and call reports. These noninterest expenses typically equal 4–6 percent of outstanding balances for large issuers and are mostly (but not entirely) a proper component of total cost. The exception is that the expense of generating a new account (mostly advertising and marketing costs) should properly be considered an investment and thus should be amortized over a longer period. Nevertheless, I have no systematic way to separate out these new-account expenses from banks' profits; consequently, I use the entire "noninterest expense" in my computations. Observe that this will tend to overstate costs and understate the returns on assets and equity. Loan losses are best measured by the bank's "net credit losses" or "net charge-offs," which represent the outstanding balances that the bank newly treats as uncollectable.²¹ (Typically, a bank charges

off a balance six months after the cardholder ceases payment on an account.) MBNA's charge-off rate is 1–2 percent lower than that of many large credit card issuers.

As seen in Table 5, MBNA's *pretax* return on assets (ROA) in credit cards equaled 7.09 percent in 1985, 6.63 percent in 1986, and 4.80 percent in 1987. The 1987 figure, for example, interacts with a 42-percent tax rate to yield an *after-tax* return on assets of 2.78 percent. For evidence of the accuracy of this computation, one need look no further than MNC Financial's 1987 annual report:

Our credit card operations had another outstanding year. Maryland Bank, N.A. (MBNA) is by no means typical of the industry, which often is the target for criticism and concern. Over the past five years, MBNA has been one of our fastest growing businesses. With \$2 billion in outstandings, it continues to be a low cost, high-volume producer with chargeoffs of about 2%—about half of the industry average. We think most investors will find it hard not to be impressed with a business that earns more than 2.5% (after-tax, 1987) on assets.²²

By way of comparison, the bank holding company as a whole earned a 1.36-percent ROA before taxes and a 1.00-percent ROA after taxes in 1987.²³ The holding company, minus its credit card business, earned less than a 0.80-percent ROA after taxes in 1987.

B. *The Ordinary Rate of Return in the Banking Industry*

The pretax return on assets for all U.S. commercial banks during the sample years

²¹The follow-up bank credit card survey specifically asked for the bank's "net credit losses" (see Appendix A). The item used from call reports is the "net charge-offs." An alternative measure of losses that could have been used from the call reports is the bank's "provision for loan losses," which is often higher and which may include an allowance for loans that the bank (statistically) expects to charge off in the future. There are two reasons not to use the figure for provision. First, credit card accounts incur most of their charge-offs in the initial two years of the life of the account. Hence, the difference between "provision" and "net charge-offs" (and, in fact, some of "net charge-offs" itself) typically represents an expense of generating new accounts and, as in the case of marketing expenses, should properly be treated as an investment which is amortized over a longer period. Using "net charge-offs" mitigates this effect and gives a better measure of cost. Second, "provision" is a quantity that is easily manipulated by the bank: one can use a large loss provision to defer income taxes or a small

loss provision to report high current earnings. "Net charge-offs" is less manipulable. The Federal Reserve System's functional cost analysis also uses "net credit losses" in earnings computations.

²²MNC Financial, 1987 Annual Report (dated March 1988), p. 4.

²³MNC Financial, 1987 Annual Report (dated March 1988), p. 1.

equaled 0.85 percent in 1983, 0.83 percent in 1984, 0.90 percent in 1985, 0.80 percent in 1986, 0.28 percent in 1987, and 1.14 percent in 1988.²⁴ Taking into account that some areas of banking were effectively taxed at a lower rate than the credit card businesses (which were taxed at close to the statutory tax rates of 34–46 percent during this period), it is probably correct to think of 1.20 percent as the ordinary (pretax) return on assets in the banking industry at large.

The relationship between the ordinary rate of return on assets and the ordinary rate of return on equity in the banking industry depends on the capital requirements of banks. For the period 1983–1988, the capital requirement equaled about 6 percent of assets. First, in 1984–1985, U.S. banking regulators promulgated capital standards for all commercial bank activities equaling 6 percent of assets for total capital (and 5.5 percent of assets for primary capital).²⁵ Second, *actual* total equity capital for all insured U.S. commercial banks equaled 5.96 percent of assets in 1983, 6.01 percent in 1984, 6.17 percent in 1985, 6.21 percent in 1986, 6.06 percent in 1987, and 6.10 percent in 1988 (with substantially smaller percentages for the larger banks).²⁶ Dividing an ordinary (pretax) return on assets of 1.20 percent by a capital requirement of 6 per-

cent would imply an ordinary (pretax) return on equity of 20 percent per year.

C. Computations of Ex Post Profitability for 15 Large Issuers

Several different summary measures of (*ex post*) profitability are presented in Tables 6 and 7. The first measure of return on assets, ROA (reported), is precisely the calculation we illustrated above for MBNA. One potential flaw in this calculation is that it relies on the bank's own reported cost of funds. The problem here is that some banks may not have been allocating the true opportunity cost of their low-cost core deposits (e.g., passbook accounts and non-interest-bearing checking accounts) to their credit card businesses; in that event, some of the profits allocated to the credit card operations would in fact be attributable to the branch banking business.

This difficulty is easily remedied by replacing each bank's reported interest expense with the standardized index of the cost of funds defined and defended in Section II. My second measure of return on assets, ROA (adjusted), is computed by using an interest expense of COST OF FUNDS applied to the nonequity portion of assets; thus, interest expense as a percentage of assets equals 94 percent of COST OF FUNDS.²⁷ If anything, my adjustment tends to reduce systematically the reported returns; observe in Table 7 that ROA (reported) exceeds ROA (adjusted) in four out of six years.

The return on equity is computed in two different ways in Table 7. The first and most obvious measure, ROE (actual cap), merely divides (pretax) profits by the actual capital residing in the credit card bank at the previous year's end. (Since each bank in Table 7 is a legally distinct entity, its capital is a

²⁴*Federal Reserve Bulletin*, July 1989, p. 462 (table 1), and July 1988, p. 404 (table 1). The substantially lower earnings for 1987 reflect the decision of large banks to set aside large sums to cover troubled loans to developing countries. Excluding international operations, the banks' rates of earnings in 1987 appear to have very slightly exceeded those of 1986.

²⁵At this writing, bank capital standards are scheduled to rise, by international agreement, to 8 percent of total risk assets in 1992. At the same time, the new and rapidly expanding practice of securitizing credit card assets has the effect of removing the credit card accounts from banks' balance sheets, thus reducing the effective capital requirement.

²⁶*Federal Reserve Bulletin*, July 1989, pp. 474–83 (table A.1), and July 1988, p. 405 (table 2). For money-center banks, equity capital equaled 4.30, 4.56, 4.69, 4.78, 4.33, and 4.42 percent of assets in the respective years. For other banks with \$5 billion or more in assets, equity capital equaled 4.76, 5.08, 5.42, 5.50, 5.29, and 5.29 percent of assets in the respective years.

²⁷ROA (adjusted), as computed from the call report data, also contains a second, minor adjustment: to the extent that a bank has purchased credit card portfolios from other banks at a premium (see Section V) and subtracted a portion of the premium from its profits, ROA (adjusted) adds it back in.

TABLE 6—RETURN ON ASSETS AND RETURN ON EQUITY (PRETAX) BASED ON DIRECT REPORTS OF CREDIT CARD ISSUERS

			Percentage returns				
Bank	Rank	Measure	1984	1985	1986	1987	1988
A	1-10	ROA (reported)	5.98	4.36	5.59	6.91	N.A.
		ROA (adjusted)	6.09	6.06	6.64	6.53	N.A.
		ROE (adjusted)	101.5	101.1	110.7	108.8	N.A.
B	11-20	ROA (reported)	8.09	7.78	3.74	4.21	N.A.
		ROA (adjusted)	8.00	7.22	7.15	5.04	N.A.
		ROE (adjusted)	133.3	120.3	119.1	84.0	N.A.
C	11-20	ROA (reported)	5.01	5.70	3.82	4.55	4.30
		ROA (adjusted)	5.00	6.21	6.49	5.33	4.23
		ROE (adjusted)	83.3	103.5	108.2	88.8	70.5
D	21-30	ROA (reported)	7.20	7.86	7.81	7.93	8.05
		ROA (adjusted)	3.28	6.39	7.33	6.39	5.84
		ROE (adjusted)	54.7	106.5	122.2	106.5	97.4
E	21-30	ROA (reported)	8.48	8.92	8.74	9.96	9.69
		ROA (adjusted)	7.48	9.20	10.02	8.09	7.31
		ROE (adjusted)	124.6	153.3	166.9	134.8	121.8
F	31-40	ROA (reported)	7.27	7.54	6.41	8.39	5.87
		ROA (adjusted)	5.75	6.62	5.91	6.98	4.95
		ROE (adjusted)	95.9	110.3	98.5	116.4	82.5
G	41-50	ROA (reported)	2.24	6.37	6.26	6.15	2.75
		ROA (adjusted)	1.67	6.54	6.41	5.81	2.74
		ROE (adjusted)	27.8	109.0	106.8	96.9	45.6
Direct report averages:		ROA (reported)	6.32	6.93	6.05	6.87	6.13
		ROA (adjusted)	5.32	6.89	7.14	6.31	5.01
		ROE (adjusted)	88.7	114.9	118.9	105.2	83.6

Source: Author's bank credit card survey (Appendix A, follow-up survey, Table A2, question 4).

well-defined quantity.²⁸) However, ROE (actual cap) is not an entirely appealing measure of return on equity. Observe that, for example, the quantity of capital that resides in Citibank South Dakota (as opposed to the principal New York bank or the parent holding company) is relatively discretionary and arbitrary. Indeed, one finds that the credit card subsidiary of a bank is often relatively undercapitalized in

some years and relatively overcapitalized in other years. While ROE (actual cap) is reported in Table 7, this datum should probably be interpreted skeptically.

A preferred measure to consult is ROE (adjusted), which is computed simply by dividing ROA (adjusted) by 6 percent. The logic behind this measure is that, as argued above, the capital requirement during the sample period has in practice equaled about 6 percent of assets, uniformly across banking activities. Thus, it seems more sensible to impute the 6-percent capital standard to all credit card assets than to rely on a capriciously chosen bank number. ROE (adjusted), which is the last measure provided in Tables 6 and 7, is probably the most informative to examine and discuss.

²⁸Data on credit card capitalization for the firms reported in Table 6 do not exist. In fact, six of the seven firms operate their credit card businesses within the same bank as their other lines of business, so there does not exist capital separately allocated to the credit card business; "actual capital," then, is not a well-defined quantity.

TABLE 7—RETURN ON ASSETS AND RETURN ON EQUITY (PRETAX) BASED ON CALL REPORTS FILED WITH THE FDIC

Bank	Rank	Measure	Percentage returns					
			1983	1984	1985	1986	1987	1988
Citibank (South Dakota), N.A.	1	ROA (reported)	7.44	7.05	6.02	6.24	5.31	4.26
		ROE (actual cap)	81.4	80.3	75.5	75.1	77.7	48.6
		ROA (adjusted)	5.40	5.24	5.56	6.62	5.61	3.92
		ROE (adjusted)	90.1	87.4	92.7	110.3	93.5	65.4
Chase Manhattan Bank (U.S.A.)	2	ROA (reported)	4.73	5.63	5.29	6.03	4.50	3.84
		ROE (actual cap)	56.4	71.2	109.1	146.5	75.2	53.9
		ROA (adjusted)	3.32	3.62	4.28	5.46	3.24	2.68
		ROE (adjusted)	55.3	60.4	71.3	91.1	54.0	44.7
Maryland Bank, N.A.	7	ROA (reported)	8.14	7.35	7.09	6.63	4.80	N.A.
		ROE (actual cap)	74.8	142.0	121.6	128.8	86.8	N.A.
		ROA (adjusted)	7.72	6.76	8.04	7.67	4.86	N.A.
		ROE (adjusted)	128.6	112.7	134.0	127.8	81.0	N.A.
Beneficial National Bank (U.S.A.)	15 ^a	ROA (reported)	4.23	5.10	3.20	1.61	N.A.	N.A.
		ROE (actual cap)	34.3	77.7	39.6	22.0	N.A.	N.A.
		ROA (adjusted)	1.31	4.01	4.27	3.70	N.A.	N.A.
		ROE (adjusted)	21.9	66.9	71.1	61.7	N.A.	N.A.
Lomas Bank U.S.A.	18	ROA (reported)	N.A.	N.A.	N.A.	3.59	4.74	4.35
		ROE (actual cap)	N.A.	N.A.	N.A.	60.6	80.7	39.9
		ROA (adjusted)	N.A.	N.A.	N.A.	4.47	5.20	4.71
		ROE (adjusted)	N.A.	N.A.	N.A.	74.5	86.6	78.5
CoreStates Bank of Delaware	20	ROA (reported)	2.86	4.21	5.09	6.46	5.03	4.10
		ROE (actual cap)	63.9	61.9	100.1	131.8	95.1	70.1
		ROA (adjusted)	2.71	4.65	6.14	7.55	5.91	4.57
		ROE (adjusted)	45.1	77.6	102.3	125.9	98.6	76.2
First City Bank—Sioux Falls	21	ROA (reported)	N.A.	N.A.	6.88	4.15	5.34	6.85
		ROE (actual cap)	N.A.	N.A.	74.0	41.6	43.4	48.5
		ROA (adjusted)	N.A.	N.A.	5.82	4.56	5.59	6.58
		ROE (adjusted)	N.A.	N.A.	97.1	75.9	93.2	109.6
First Omni Bank, N.A.	33	ROA (reported)	8.68	5.98	8.07	6.77	5.69	5.01
		ROE (actual cap)	29.8	32.2	58.0	99.8	93.2	72.4
		ROA (adjusted)	6.19	3.36	6.51	6.85	5.51	4.38
		ROE (adjusted)	103.2	56.0	108.5	114.2	91.8	73.0
Avco National Bank	44 ^a	ROA (reported)	N.A.	4.82	3.72	1.19	N.A.	N.A.
		ROE (actual cap)	N.A.	344.6	103.6	27.9	N.A.	N.A.
		ROA (adjusted)	N.A.	4.23	4.66	5.06	N.A.	N.A.
		ROE (adjusted)	N.A.	70.5	77.7	84.3	N.A.	N.A.
Call-report averages:		ROA (reported)	6.01	5.73	5.67	4.74	5.06	4.74
		ROE (actual cap)	56.8	115.7	85.2	81.6	78.9	55.6
		ROA (adjusted)	4.44	4.55	5.66	5.77	5.13	4.47
		ROE (adjusted)	74.0	75.9	94.3	96.2	85.5	74.6

Source: Consolidated reports of condition and income (call reports).

^aBank's credit card portfolio was acquired in 1986–1987; see Table 9.

By the standards of the previous subsection, the rates of return reported in Tables 6 and 7 are extraordinary. All seven banks that provided direct reports of profit data for 1985 and six out of seven that provided direct reports for 1986 attained (pretax) returns on equity exceeding 100 percent per year! The sample average for these banks also exceeded 100 percent in 1987 and exceeded 80 percent in 1984 and 1988. The profit figures drawn from call reports are not quite as large but still are generous. Sample averages for return on equity exceeded 90 percent in 1985 and 1986 and exceeded 70 percent in all remaining years. It is unclear why the direct reports provided returns on equity systematically 10–20 percentage points higher than the call reports. Part of the reason is undoubtedly that firms that established separate credit card banks sustained higher rates of growth in assets, concealing a greater investment in new accounts in the cost data.²⁹

D. *Estimates of Ex Post Profitability for an Additional Eight Issuers*

The computations of the preceding subsection were performed for 15 credit card issuers for whom all components of profits were known. If the data set is expanded to include banks for whom most, but not all, components are known, it is possible to make statements about an even larger proportion of the 50 largest firms.

Eight additional banks, including Bank of America (ranked third largest), First Chicago (ranked fourth), and Manufacturers Hanover (ranked fifth), have disclosed significant information in connection with the issuance of credit-card-backed securi-

ties.³⁰ All items in Table 5 except noninterest expenses (i.e., non-credit-related operating expenses) and interchange fees are thus known for these issuers; moreover, these two missing items (unlike customer revenue and net charge-offs) do not vary widely among comparable banks. I approximate their rates of profit by assuming that the additional eight issuers' operating expenses and interchange fees are equal (as a percentage of assets) to those for which I have direct knowledge. Making the same normalizations for these banks as before, I obtain conservative³¹ average pretax returns on equity (adjusted) of 65 percent in 1984 (three banks), 87 percent in 1985 (five banks), 92 percent in 1986 (all eight banks), 76 percent in 1987 (seven banks), and 62 percent in 1988 (six banks). It is worth noting that all of these numbers are quite close to the adjusted ROE's of Table 7, and all exceed 60 percent per year.

The sample from which a conclusion about profitability can be based is rather large. Exact computations or good estimates are available for as many as 23 of the 50 largest issuers of bank credit cards, with approximately 50 percent of industry market share (by outstanding balances). Included in the sample are all of the five largest firms.

The conclusion drawn from Table 6, Table 7, and the numbers stated two paragraphs above is quite straightforward. As argued above, the ordinary (pretax) return on equity in banking is on the order of 20 percent per year. Credit card businesses earned annual returns of 60–100 percent or more during the years 1983–1988. Plastic earned

²⁹An anonymous referee correctly observed that the return on assets derived from direct reports uses credit card balances in place of total assets as denominator. However, the call report data show that this does not appreciably overstate the true return on assets. Premises and fixed assets typically equaled no more than 0.3–2.0 percent of a credit card bank's total assets, while credit card loans generally exceeded 97 percent of total assets.

³⁰See Appendix B for a listing of the relevant registration statements. While information in connection with credit-card-backed securities is also available for Citibank, Maryland National, and Lomas Bank U.S.A., these banks are already represented in Table 7 and so are excluded from the current discussion.

³¹Some prospectuses report gross rather than net charge-offs, reducing the reported level of profits. Furthermore, the figures I used for noninterest expense (generally 5.40 percent of outstanding balances) seem to be on the high side.

strongly positive economic profits: the credit card business earned 3–5 times the ordinary rate of return in the banking industry.

IV. Additional Evidence from the Functional Cost Analysis

The previous discussion has focused on the profitability of the 50 largest issuers of credit cards during the period 1983–1988. It is interesting to consider briefly whether the conclusions change when the sample of banks and the time period examined are extended.

This exercise should help address two potential questions.³² First, suppose it were the case that the smaller players in the market earned only the ordinary rate of return on capital. Then, the reader may be troubled by the possibility that the larger firms may possess some unobservable attributes which are not reproducible (e.g., “business acumen”) that earn positive rents. (However, there would still remain the question of why 50 larger firms is not enough for competition.) On the other hand, if literally hundreds of firms, including small regional banks, all earn supranormal profits, it becomes much more convincing that none of these firms possesses anything special (except for a base of customers). Second, the reader may be concerned that the article has focused on a period of time that coincides with a cyclical boom in the national economy. Part of this selection of time is unavoidable: as observed in Section I, the credit card market did not become functionally deregulated until about 1982; much of the earlier experience can be dismissed as the result of state usury ceilings. In addition, my reliable sources of data only extend back to this time. Nevertheless, it may be helpful to present some (albeit imperfect) data which provide a better sense of the extent to which profitability is cyclical.

Both of these points can be discussed by introducing an additional source of data: the Federal Reserve System’s functional cost analysis (FCA). In this author’s opinion, the FCA data are considerably less reliable than the other profitability data utilized in this article, and so they should be interpreted cautiously. There has been little effort to track the same banks from year to year; in particular, the sample size has dropped approximately 60 percent from 1976 to 1988. One also obtains the sense that the accounting data provided by the smaller banks in the Fed’s sample do not do as good a job as the other sources of properly allocating costs between credit cards and the banks’ other lines of business. Nevertheless, it is the only available source of profitability figures for the smaller banks and for earlier years.

The average charge-off rate and the average return on assets from the functional cost analysis for the years 1976–1988 are displayed in Table 8. The typical credit card issuer represented in the FCA sample is ranked approximately between number 300 and 400, nationally, by outstanding balances. The first observation to make is that the charge-off experience of the smaller banks is broadly consistent with what has already been seen in Table 4. As before, loan losses remained tightly in the 1–3 percent per year range.

Since bank credit card operating expenses are believed to exhibit increasing returns to scale over a range (Christine Pavel and Paula Binkley, 1987), one would expect that the return on assets would be somewhat lower than that reported above for the largest issuers. Indeed, during the period 1983–1988 (for which numbers are available for both the smaller issuers and the larger issuers), the smaller banks earned only about 60 percent of the returns of the larger banks and only about 50 percent of the “excess” returns of the larger banks. Nevertheless, the returns do remain substantially above the ordinary rate of return in banking; over the longer period that includes the cyclical downturn, the smaller issuers still appear to have earned roughly twice the ordinary rate of return in banking.

³²These are questions that were raised by two anonymous referees. I am grateful to them for raising these issues.

TABLE 8—RETURN ON ASSETS FOR SMALLER CREDIT CARD ISSUERS, 1976–1987, FROM THE FEDERAL RESERVE SYSTEM'S FUNCTIONAL COST ANALYSIS (FCA)

Year	Number of banks in sample	Average charge-off rate (percentage)	Average return on assets (percentage)
1976	236	1.48	2.72
1977	224	1.41	3.07
1978	181	1.66	2.44
1979	184	1.92	1.60
1980	139	2.54	-1.52
1981	128	2.25	0.97
1982	138	1.93	2.40
1983	102	1.58	2.37
1984	98	1.24	3.45
1985	85	1.81	3.97
1986	76	2.33	3.28
1987	93	1.65	3.94
1988	89	2.51	2.72

Notes: Data are taken from the Board of Governors of the Federal Reserve System's *Functional Cost Analysis, National Average Report, Commercial Banks, Credit Card Function (Card Banks), 1976–1988*. The third column represents "net credit losses + net fraud losses"; the fourth column represents "net earnings after cost of money." Both columns are expressed as percentages of average outstanding balances on credit cards and are weighted averages based on individual banks' average outstanding balances. Earnings are before taxes. In Section III-B, it is argued that the ordinary return on assets is approximately 1.20 percent.

These numbers may significantly understate the profitability of credit cards over the entire business cycle. First, banks have the ability (with some lags) to increase their lendings at cyclical peaks and to cut their lendings at cyclical troughs. (This is not merely conjectural: their ability to act in this way is fairly apparent from data on banks' levels of outstanding balances.) Thus, greater weight should be placed on the returns in boom years. Second, as emphasized above, banks took some time to learn how to exploit the December 1978 *Marquette* decision and so were hindered in their ability to react to climbing costs of funds during 1979–1981. The learning has now been done, and so the 1979–1981 earnings experience is unlikely to be repeated.

V. The *Ex Ante* Profitability of the Credit Card Market

As was emphasized in the fourth paragraph of Section III, there is good reason to be a bit skeptical of *ex post* profitability data. In this section, I will seek (as directly as possible) to examine *ex ante* returns of the credit card market.

Suppose that bank I has issued credit cards to consumers and has \$*X* in balances outstanding on these accounts. The question one may ask is how much bank II will pay bank I to acquire these accounts, as a function of *X*. If the credit card accounts were expected, *ex ante*, to pay the ordinary rate of return on capital (or the risky equivalent), then the transaction would presumably occur at about *par* (i.e., bank II would pay bank I the same \$*X* to assume the accounts). If there existed a substantial probability that consumers would default on the loans and if the contractual interest payments and fees did not adequately compensate for this eventuality, then bank II could presumably acquire these loans at a *discount*.³³ Finally, only if the owner of the credit card accounts could be expected to earn above the ordinary rate of return on capital would the accounts sell at a *premium* above \$*X*; then, the future stream of positive economic profits would be capitalized in the transaction price. If bank II pays bank I \$120 million for accounts on which only \$100 million is owed, I will refer to this as a 20-percent premium.

The model of perfect competition predicts that resales of credit card accounts will occur at *par*, in the long run.³⁴ During the

³³For a good example of discounted loans, consider the resale among banks of loans to developing countries; in the late 1980's, such transactions frequently occurred at prices well below 50 percent of face value.

³⁴One may argue that there should exist a premium representing the cost of establishing an ongoing business. An important point to observe is that, typically, when a bank acquires another bank's credit card portfolio, it transfers the acquired portfolio over to its own preexisting offices and processing facilities. That is to say, basically the only portion of the "ongoing business" that the acquirer desires is the customer base. In the

TABLE 9—PREMIA IN REALES OF CREDIT CARD ACCOUNTS

Date	Seller	Buyer	Outstanding balances	Premium reported (percentage)	Primary source
April 1984	Continental Bank	Chemical Bank	\$824 million	21	WS April 2, 1984
December 1986	Texas American Bancshares	Republic Bank	\$50 million	14	WS January 2, 1987
December 1986	Beneficial National Bank	First Chicago	\$1,100 million	13	AB January 2, 1987
February 1987	National Bancshares Texas	Lomas & Nettleton	\$41 million	14	WS February 23, 1987
April 1987	Louisiana Bancshares	Lomas & Nettleton	\$157 million	16	WS April 16, 1987
May 1987	Avco National Bank	Household Bank	\$322 million	19	WS April 29, 1987
July 1987	Bank of Mid-America	Lomas & Nettleton	\$120 million	19	WS July 23, 1987
July 1987	Colonial National Bank	Household Bank	\$317 million	11	WS July 23, 1987
September 1988	First Republic Bank Delaware	Citicorp	\$623 million	25	NY September 12, 1988
November 1988	Equibank	CoreStates Bank	\$100 million	25	KP
February 1989	Meritor Financial Group	Chase Manhattan	\$85 million	24	AB February 2, 1989
March 1989	Society for Savings Bancorp	First Chicago	\$230 million	18	NY April 13, 1989
May 1989	Michigan National Bank	Chase Manhattan	\$1,100 million	21	WS July 19, 1989
May 1989	Empire of America	Citicorp	\$650 million	3	AB June 2, 1989
June 1989	Colonial National Bank	Household Bank	\$98 million	25	AB June 23, 1989
July 1989	Leader Federal Savings	Chase Manhattan	\$36 million	20	KP
August 1989	California Federal Bank	Household Bank	\$125 million	18	AB September 18, 1989
September 1989	Chevy Chase Savings	CoreStates Bank	\$200 million	23	AB September 18, 1989
September 1989	Imperial Savings & Loan	Wells Fargo Bank	\$280 million	22	WS January 11, 1990
September 1989	Dreyfus Corp.	Bank of New York	\$790 million	21	AB September 20, 1989
October 1989	First City Bancorporation	Bank of New York	\$552 million	24	AB October 19, 1989
December 1989	Bank South	Society National	\$41 million	24	AB December 27, 1989
December 1989	Bank of Boston	Chase Manhattan	\$625 million	23	AB January 5, 1990
December 1989	Investors Savings Bank	Chase Manhattan	\$24 million	25	AB January 5, 1990
January 1990	Bank of New England	Citicorp	\$652 million	27	NY January 30, 1990
February 1990	Colonial National Bank	Household Bank	\$50 million	20	AB March 7, 1990
April 1990	Fleet/Norstar	Norwest	\$200 million	20	AB April 3, 1990

Sources: WS = *Wall Street Journal*; NY = *The New York Times*; AB = *The American Banker*; KP = Kidder Peabody (Anderson and Deans, 1989).

equilibrating process toward the long run, the theory would tolerate discrepancies from par but firmly predicts that resale prices will monotonically converge toward par. However, a systematic failure of competition in the credit card market (as suggested by the profit figures in the previous sections) would require, to the contrary, that interbank transactions persistently occur at substantial premia. Fortunately, there exist real data against which to test these two divergent predictions.

The premia paid by banks in credit card deals during the years between 1984 and

early 1990 are compiled in Table 9. All 27 such interbank transactions for which I could find a public disclosure of the premium are reported. The average premium in Table 9 equals 20 percent; all transactions occurred at premia between 3 percent and 27 percent. There is no tendency for the premia to vanish monotonically; if anything, the largest premia are associated with the most recent transactions.

The resale data clearly suggest that, at this writing in 1990, and throughout the period of 1984–1988, the *ex ante* expected economic profits (adjusted for risk) on existing credit card accounts were substantially positive. The premium that is theoretically justified for credit card accounts depends on a number of parameters for which I possess no data. However, Table 10 presents the premia that are justified for a

model of perfect competition, customers inexorably gravitate to the low-priced firm; the phenomenon of “captive” or “loyal” customers does not exist. Thus, an existing base of customers, by itself, should draw no premium.

TABLE 10—PREMIA IN RESALE OF CREDIT CARD ACCOUNTS JUSTIFIED BY VARIOUS EXPECTED LIFETIMES OF ACCOUNTS, ANNUAL GROWTH RATES OF OUTSTANDINGS PER ACCOUNT, AND CREDIT CARD PROFITABILITY (AS A MULTIPLE OF ORDINARY RATE OF RETURN)

Expected lifetime (years)	Multiple of ordinary rate of return	Premia (percentage)			
		10-percent growth	20-percent growth	30-percent growth	40-percent growth
2	1	0.00	0.00	0.00	0.00
	2	2.40	2.51	2.62	2.73
	3	4.80	5.02	5.24	5.45
	4	7.20	7.53	7.85	8.18
	5	9.60	10.04	10.47	10.91
4	1	0.00	0.00	0.00	0.00
	2	4.80	5.50	6.27	7.15
	3	9.60	10.99	12.55	14.29
	4	14.40	16.49	18.82	21.44
	5	19.20	21.98	25.10	28.58
6	1	0.00	0.00	0.00	0.00
	2	7.20	9.05	11.38	14.30
	3	14.40	18.10	22.76	28.60
	4	21.60	27.15	34.15	42.90
	5	28.80	36.19	45.53	57.20

Calculations: For N = expected lifetime of credit card accounts, g = annual growth rate of outstandings per account (during lifetime of account), $r = 0.10$ = interest rate used by bank in discounting, ROA_{cc} = return on assets from credit cards, $ROA_{ord} = 0.012$ = ordinary return on assets in banking, and Φ = premium in resale of credit card accounts (as proportion of outstanding balances at time of sale),

$$\Phi = (ROA_{cc} - ROA_{ord}) \sum_{K=0}^{N-1} \left(\frac{1+g}{1+r} \right)^K$$

number of sets of assumptions. The expected "lifetime" of an account represents the number of years that a bank anticipates that the consumer will continue to maintain his credit card account, under the same borrowing patterns and the same rate of profitability. The growth rate represents the rate at which a bank believes that the outstanding balances on the acquired credit card accounts will increase during their lifetime; meanwhile, I assume that the bank discounts using a 10-percent interest rate. The ordinary rate of return in banking is taken, as above, to be a 1.20-percent annual pretax return on assets. The calculations in Table 10 implicitly assume that the seller of credit card accounts receives all of the gains from trade; if (as one may reasonably expect) the buyer also obtains some gains, then the indicated premia require still higher rates of return from credit cards. Thus, these should only be taken as lower bounds on implied profitability.

As Table 10 indicates, it is difficult to justify the recent flurry of premia in the range of 23–27 percent unless returns equaling at least three times the ordinary rate of return in banking are expected to persist. Even with the optimistic³⁵ assumptions that the typical account will be maintained for six years after the acquisition and that the outstanding balances will grow at an overly fast 40 percent per year during that period, profits at three times the ordinary rate of return lead to a premium of only 28.60 percent. The more typical premium of 20 percent, in conjunction with reasonable projections of lifetime and

³⁵For example, at the date of the First RepublicBank transaction, First Republic's average outstanding balance per active account equaled about \$1,000, roughly the national average. A 40-percent growth rate for five years would increase the average outstanding balance by more than a factor of five, justifying the adjective "optimistic."

growth, still requires profitability of three or more times the ordinary rate of return in banking. Finally, it should be observed that one of the lowest reported premia (Colonial National Bank, at 11 percent) involved credit cards with a mean outstanding balance per active account equaling \$2,000, or about twice the national average. It would be rather unrealistic to project growth of more than 10 percent per year on these balances, which still suggests profitability about three times the ordinary rate.

I will make three final notes on *ex ante* profitability. First, the reader may still worry that, since credit card debts are unsecured, charge-off rates will jump and profitability will plummet in the next recession. Some historical data should allay these concerns. In Tables 4 and 8, one may trace back net charge-offs through the last recession. One finds, for these samples, that charge-offs in the early 1980's increased only fractionally above prior years and peaked at only about 2.5 percent of outstandings. Independently, Visa system-wide data traces back credit losses through the last two recessions, and finds (annualized) quarterly charge-off rates peaking at 3 percent in 1974 and again in 1980.³⁶ Solicitation of new accounts, and not cyclical phenomena, are the important contributors to credit card charge-offs (see also Section II-B, above).

Second, the reader may have noted that substantial premia (although not as large as for credit card accounts) have also been reported in sales of regional banks. If anything, this makes the credit card premia even more surprising. The premia for regional banks represent "goodwill," which economists should interpret as economic rents derived from local monopolies in banking. By way of contrast, the national market for credit cards has no local monopolies, so the competitive model predicts that "goodwill" should equal zero.³⁷

Third, the magnitude of resale premia may be taken as clear evidence that players within the industry itself attach little credence to the possibility that the credit card market will begin to behave competitively in the years immediately following 1990. (If, for example, it were believed that competition would drive economic profits to zero in two years, this would be the same as using an expected "lifetime" in Table 10 of two years.) Data in the previous section showed that the zero-profit prediction failed in the years 1983–1988. If bankers have had rational expectations in their acquisitions of accounts, then supranormal profits should persist for at least the period 1990–1993. Credit card profits will then have equaled three times the ordinary rate of return for more than an entire decade, certainly an extended adjustment period to the long run!

VI. Theoretical Explanations for the Failure of Competition

In this section, I briefly outline some theoretical models that lead to predictions of price stickiness and positive economic profits. I also discuss some empirical evidence related to these theories. Formal modeling details are available in Ausubel (1988), the working-paper precursor of this article.

A. Search/Switch Cost Theories

One of the common explanations offered for high credit card interest rates is that consumers find it difficult to locate banks offering favorable terms. Indeed, the U.S. Congress enacted legislation in October 1988 that requires all issuers to disclose their interest rates, fees, and grace periods on solicitations and applications; supporters of the bill articulated the rationale of enabling consumers to shop around for the least expensive card, (i.e., of reducing consumer search costs).

Models with search costs may plausibly lead to sticky interest rates and positive

³⁶The source of system-wide net charge-offs is Standard & Poor's *Asset-Backed Securitization CreditReview*, March 16, 1987, p. 19.

³⁷In addition, a regional bank may own appreciated real estate whose book value equals historic cost, also contributing to reported premia over book value. The

credit card transactions typically do not involve any real assets that have appreciated in value.

profits. While other explanations also exist for price stickiness (e.g., menu costs; see e.g., Julio J. Rotemberg and Garth Saloner, 1987), there is good reason to focus on search/switch costs in the credit card market. Banks have recently begun to use marketing techniques that are consistent with this type of story. Issuers frequently waive the annual fee for a fixed period of time and, in a few cases, offer "to pay you up to \$100 when you transfer your other credit card balances" to their MasterCard or Visa accounts.³⁸ The focus of federal regulation on disclosure provides additional support for the search-cost explanation.

In models that are thematically related to that of Peter Diamond (1971) and subsequent papers, there may exist a continuum of symmetric equilibrium prices that are consistent with any given marginal cost.³⁹ Therefore, the historical price may continue to be an equilibrium even after a change in marginal cost. That is, price stickiness may be consistent with equilibrium. (Detailed analysis of a straightforward model that yields this conclusion is provided in the working-paper version of this paper. It should also be noted that a formal model of switch costs [see e.g., Ausubel, 1984; Joseph Farrell and Carl Shapiro, 1987; Paul Klemperer, 1987] can result in conclusions similar to those from search-cost models.)

Such models also provide a reason why supranormal profits may not be competed away. If prices remain sticky when costs drop, firms begin to earn supranormal profits. Profits continue at high levels until prices unstick, costs rise again, or the customer base is sufficiently eroded. The logic is that, if consumers face search costs in locating (or face switch costs in moving to) lower-priced firms, then higher-priced firms can hold onto many of their (captive) customers despite their high prices. As suggested

above, competitors may try to defeat this inertia by offering sign-up bonuses to new customers, but to the extent that such devices are limited in their effectiveness and practicality, firms may derive supranormal profits from their existing customer base. Finally, observe that this story enables a firm with a base of "loyal customers" to earn supranormal profits despite competition both from other existing firms (who want to increase their own customer bases) and from new entrants (who want to establish customer bases).

The credit card industry is a business where both search costs and switch costs are likely to be especially prevalent. They include: (a) the information cost of discovering which banks are offering lower interest rates; (b) the cost in time, effort, and emotional energy in filling out an application for a new card (and possibly getting rejected); (c) the fact that the card fee is usually billed on an annual basis, so that if one switches banks at the wrong time, one forgoes some money; (d) the perception that one acquires a better credit rating or a higher credit limit by holding the same bank's card for a long time; and (e) the time lag between applying for a card and receiving one.

While credit card consumers undoubtedly face some positive level of search costs and switch costs (and this gives entirely rational justification for the observed market behavior), there remains an empirical question as to whether the actual search/switch costs are of sufficient magnitude to justify what is observed. A typical credit card account in the late 1980's had an outstanding balance slightly over \$1,000 (see Section VI-C and Table 11, below). The prevailing premium on resales of these accounts (see Section V and Table 9) then translates to almost \$250 per account. In a search/switch cost equilibrium, one would expect the resale premium to equal the search or switch cost; yet it is hard to imagine that the costs enumerated in the previous paragraph are the monetary equivalent of \$250! Given this caveat, it is not at all clear that search or switch costs could provide a full explanation of observed market behavior; it would be valuable for future empirical work (using data at

³⁸ The quotation is taken from a direct-mailed solicitation, dated April 1989, from Imperial Savings (approximately the fifth-largest S&L issuer of credit cards).

³⁹ Mitchell Berlin and Loretta J. Mester (1988) tested and rejected a (very different) model, in which consumer search costs were used to try to explain price dispersion in credit card interest rates.

TABLE 11—PERCENTAGE OF CUSTOMERS WHO AVOID FINANCE CHARGES AND AVERAGE OUTSTANDING BALANCES PER ACTIVE ACCOUNT

Year	Percentage of customers who avoid finance charges	Number of banks reporting percentage of customers	Average outstanding balance of active account	Number of banks reporting outstanding balance
1979	31.8	6	\$523	7
1980	28.7	6	\$590	8
1981	21.9	6	\$660	8
1982	21.0	6	\$726	9
1983	22.8	8	\$711	10
1984	21.9	10	\$852	10
1985	21.4	12	\$1,014	12
1986	24.6	14	\$1,018	17
1987	27.6	9	\$1,038	10

Source: Author's bank credit card survey (Appendix A, Table A1, question 8).

the customer level) to examine this question.

B. *A New Adverse-Selection Theory*

I now propose an adverse-selection theory that relies on a very specific form of irrationality (which will be given some indirect empirical support in the following subsection). Since a credit card is really quite an expensive medium on which to borrow, I posit a class of consumers who do not intend to borrow on their accounts but find themselves doing so anyway.⁴⁰ Consumers in this first class are precisely the best customers from a (rational) bank's viewpoint: they borrow at high interest rates, yet they eventually (in most cases) repay their loans. At the same time, these consumers are unlikely to be responsive to any interest rate cut by a bank, as they do not intend to borrow at the outset.

I also assume that there is a second class of consumers who fully intend to borrow on their credit card accounts. These are the consumers who are bad credit risks and thus lack less expensive alternatives; bank cards are their best sources of credit. Consumers in the second class are less than ideal from

a bank's perspective: they borrow large sums but often default. Insidiously, these customers are more likely to comparison shop on interest rates than the better credit risks, as they actually plan to be paying substantial finance charges. (There is also a third class of consumers, the "convenience" users, whom I can neglect in this discussion. They never borrow on their credit cards and, thus [rationally], are completely unresponsive to interest rate differentials.)

Given this environment of consumers, banks will be hesitant to compete in the interest-rate dimension, as a lower price on credit would disproportionately draw the class of consumers who plan to utilize their credit lines. If consumer behavior along these lines is superimposed on a search-cost model, the tendency toward interest-rate stickiness that was described in the preceding subsection becomes magnified (see Ausubel [1988], the working-paper precursor of this article, for a formalization of this story).

Such reasoning additionally provides an explanation for the apparent cross-subsidy from the transaction function to the credit function of the bank card.⁴¹ Banks only face adverse selection when they compete on the

⁴⁰It may be possible to rationalize these consumers' behavior by assuming that they face a commitment problem: consumers cannot commit their future selves not to borrow.

⁴¹Credit card issuers appear, at best, to break even on their "convenience" users and, perhaps, lose money on them. Meanwhile, issuers earn supranormal profits on consumers who borrow on their cards.

credit-sensitive portions of prices; they do not face adverse selection when they unilaterally improve the terms facing customers who charge purchases on their credit cards but do not borrow beyond the due date on their bills. This would seem to be a powerful explanation why essentially all large issuers offer a substantial grace period on new purchases (provided that the previous balance was paid in full). It also suggests why issuers hardly ever impose transaction charges, often ask for rather small (and, sometimes, zero) annual fees, and occasionally offer transaction subsidies (for example, rebates on purchases or frequent-flyer miles). At the same time, issuers may install punitive prices for bad credit risks: for example, disproportionately high fees for missing a minimum required payment. Since such large proportions of revenues are derived from finance charges, while the adverse-selection argument implies that the interest rate should not be used as an instrument for competition, it becomes much more difficult for credit card issuers to compete away profits. Thus, adverse selection helps to explain the observed extraordinary profits.

Finally, the present adverse-selection theory may be compared with that of Joseph E. Stiglitz and Andrew Weiss (1981). Stiglitz and Weiss argue that, if all banks are charging the same interest rate, no one bank will unilaterally deviate and charge a higher interest rate. The explanation is that the only consumer who would borrow at such a high interest rate is one who probably will not repay the debt (i.e., he is undertaking a very risky project). In contrast, if all banks are earning positive economic profits, the Stiglitz-Weiss effect would quicken the banks' tendencies to cut prices. A lower interest rate draws not only more customers but also better customers. Thus, Stiglitz and Weiss predict that interest rates are "upward-sticky" when costs rise and, if anything, interest rates are "downward-quick" in their model.

This is hardly a good description of real-world credit markets. Empirically, interest rates on loans have an asymmetric response to the cost of funds: they are quicker to

move upward in response to increases in the cost of funds than to move downward in response to decreases in the cost of funds. (Marcelle Arak et al. [1983] detected an asymmetric response in movements of the prime rate, and in work in progress, I have found an asymmetric response in many consumer credit markets.)

My adverse-selection theory is a *reverse* Stiglitz-Weiss effect: it creates reluctance to cut interest rates. Thus, it is a completely different and new adverse-selection theory, which may also be useful in explaining other credit markets.

C. Evidence of Consumer Irrationality

The adverse-selection theory of the previous section crucially relies on the assumption that there are consumers who do not intend to borrow but continuously do so. (Many other forms of irrationality would also render consumers insensitive to credit card interest rates.⁴²) In this subsection, I indicate some formal and anecdotal evidence of this and other forms of consumer irrationality in this market.

First, in the author's bank credit card survey, banks were asked for the percentage of their customers who pay off their full outstanding balances (and so are not subject to finance charges) and for the average outstanding balance (see question 8 in Table A1 for the exact text). The responses, summarized in Table 11, reveal that significant finance charges are being paid on the majority of credit card accounts. Despite interest rates exceeding 18 percent per year, typically *three-quarters* of active credit card accounts at major banks are incurring these high finance charges (on balances averaging over \$1,000) at any moment in time.⁴³ The proclivity of consumers to borrow at these

⁴²For example, many consumers may not understand how interest rates work and underestimate the consequences of borrowing.

⁴³The three-quarters figure should not come as a complete surprise. It would certainly have to be in this range, in order for typically 90 percent of a credit card issuer's outstanding balances to be accruing interest. See also footnote 20.

high rates suggests a substantial breakdown in optimizing behavior among credit card holders.⁴⁴ Moreover, the percentages in Table 11 are based on reliable bank data yet contradict the authoritative University of Michigan consumer survey. According to Glenn B. Canner and James T. Fergus (1987 table 3), the 1983 Michigan survey found that 47 percent of all families that use bank or retail cards "nearly always pay in full," 26 percent "sometimes pay in full," and only 27 percent "hardly ever pay in full." Unless this is evidence of a bad consumer survey, it suggests that a sizeable proportion of consumers who borrow on credit cards are unaware of how frequently they do it or, more likely, deny (to themselves and others) that they do it.⁴⁵ In this sense, the data provide indirect empirical confirmation of the presence of consumers who act as though they do not intend to borrow but who continuously do so.

Second, the experience of credit card marketers is that consumers are much more sensitive to increases in the annual fee than to commensurate increases in the interest rate, despite the fact that the majority of cardholders pay significant finance charges. This is behavior that is difficult to rationalize and is again consistent with the presence of consumers who do not intend to borrow but do so anyway.

Third, if advertising campaigns predicated on price are ineffective, it may be wondered what does attract new customers. One notable recent success has been the "Elvis card," which despite a 17.88-percent interest rate (about average) and \$36 annual fee (extremely high for a standard bank card) generated three times the response rate normally experienced by direct mail.⁴⁶

⁴⁴One would expect that optimizing behavior would lead many consumers to (a) shop around for lower-priced credit cards, (b) shift into different modes of borrowing (e.g., home equity loans), or (c) rearrange their intertemporal stream of consumption (i.e., not borrow).

⁴⁵It is possible that consumers who borrow also hold more charge accounts than those who do not borrow; but multiple accounts cannot nearly fully explain the statistical discrepancy.

⁴⁶*Credit Card News*, October 1, 1988.

Fourth, anecdotal evidence suggests that credit card consumers behave significantly different from the ideal of *Homo economicus*. This author's favorite story (heard twice, independently) involves consumers who immerse their credit cards in trays of water and place them in the freezer. The purpose of entombing the card in ice is to precommit to not making impulsive purchases.

Finally, these observations are not specifically confined to the credit card market and, in fact, are consistent with earlier work that has been done in other areas of consumer credit. One of the most surprising such articles is a study by James J. White and Frank W. Munger (1971) which found that recipients of new car loans were extremely insensitive to interest rates. It would be reasonable to expect that consumers are relatively more price sensitive in seeking out automobile loans than credit cards, as the large dollar amount would justify greater search or switch behavior. Nevertheless, White and Munger report that roughly half of the borrowers from the high-cost providers of auto loans in the Michigan locality they studied would have qualified for loans from low-cost providers. Many consumers who apparently could have borrowed at appreciably lower interest rates failed to do so. Moreover, 29 percent of the borrowers from the high-cost providers were specifically aware of at least one nearby lender who charged a lower interest rate, leading White and Munger to conclude that lack of knowledge of lower interest rates was not the principal deterrent to obtaining cheaper loans.

VII. Calculation of a "Competitive" Interest Rate

This article has thus far focused on the discrepancy between the predictions of the competitive model and actual observed behavior in the bank credit card market, while Section VIII will discuss the relative merits of regulating this market. As a bridge between these two strands of thought, this section will briefly inquire as to "competitive" interest rates: what level of interest rates would have been consistent with ordi-

TABLE 12—IMPLIED DIFFERENTIAL BETWEEN "COMPETITIVE" CREDIT CARD INTEREST RATE AND ACTUAL RATE

Year	Number of banks in sample	Adjusted return on assets (percentage)	Implied actual minus "competitive" interest rate (percentage)
1983	6	4.44	3.60
1984	13	4.94	4.16
1985	14	6.23	5.59
1986	15	6.37	5.74
1987	13	5.72	5.02
1988	10	4.72	3.91
Average:	—	5.59	4.88

Notes: Adjusted return on assets is calculated by pooling the banks in Tables 6 and 7 for each year and calculating the arithmetic average of ROA (adjusted). The number of banks in the sample reflects one overlap in the years 1984–1988.

nary returns in the credit card market in the late 1980's?⁴⁷

Suppose an explicit calculation is to be done for the year 1987. Above, I have reported the average adjusted return on assets to be 6.31 percent in the BCCS data (seven banks) and 5.13 percent in the call-report data (seven banks, with one overlap). For the following calculation, I will take 5.72 percent (the arithmetic average for the two samples) to be the actual pretax return on assets. Recall that 1.20 percent has been taken to be the ordinary pretax return on assets in the banking industry. Subtracting and taking "assets" to be equivalent to "outstanding balances," one could conclude that the excess revenues in 1987 were 4.52 percent of outstanding balances. Also recall that, typically, about 90 percent of an issuer's outstanding balances actually accrue interest. This suggests that, if interest rates had been approximately five percent-

age points lower (i.e., 4.52/0.9) in 1987, the top 50 credit card issuers would have still earned the ordinary rate of return in banking. Given that the average annual percentage interest rate for banks in this sample equaled 18.67 percent in 1987, this would imply a "competitive" interest rate of just 13.65 percent. Given that the average one-year Treasury bill yield equaled 7.52 percent in 1987, this also suggests an approximate rule of thumb that, at 1987 levels of annual fees and credit losses and with current usage patterns, the break-even point is roughly approximated by the one-year Treasury bill yield plus slightly more than six percentage points.

Analogous calculations for the period 1983–1988 are displayed in Table 12. Obviously, these calculations are sensitive to the estimate of credit card profitability. However, even using the much more conservative FCA profitability data (see Table 8), one would still find that credit card interest rates in 1987 were three percentage points above the break-even level.

VIII. Implications for Regulation

While this article has argued that the bank credit card market does not mirror the predictions of the model of perfect competition, neither does it necessarily lead to the conclusion that usury ceilings on credit card interest rates should be reestablished. As experience in many industries (e.g., airlines,

⁴⁷Obviously, this is precisely the same question that would have to be asked if the government were to choose to regulate the bank credit card market. Please note that the calculation provided here is meant only to be illustrative and would not be suitable for inclusion in any future statute without further refinement. Note that the calculation assumes, for simplicity, that credit card borrowing is perfectly inelastic in the interest rate (although this may not be a bad approximation of reality). Also note that the calculation assumes a continuous 20-percent return on equity; in fact, one would expect some degree of variation over the business cycle.

trucking, railroads, and banking itself) has demonstrated, it is often difficult to formulate a regulatory rule that unambiguously improves industrial performance. In the industry in question, the particular hazard associated with price controls is the possibility that they would impair the ability of some individuals to obtain credit cards, which are virtual necessities in certain aspects of modern life (such as renting an automobile or ordering by telephone). This section discusses the trade-offs between regulated and unregulated interest rates.

Even if this article does not criticize the recent outcome of the legislative process (i.e., rejecting the reimposition of credit card interest rate ceilings), it does at least argue that the terms of debate have been flawed.⁴⁸ Underpinning the antiregulation argument has been the market description of the credit card business (as presented in the first paragraph of this paper) and the implication that such an industrial structure inexorably leads to the perfectly competitive outcome (with all its desirable efficiency properties). For example, Martha R. Seger, a Governor of the Federal Reserve System, concluded her recent Congressional testimony on the subject by stating:

I would like to reemphasize the Board's conviction that financial markets distribute credit most efficiently and productively when interest rates are determined without artificial restraints, insofar as possible. In the credit card business, the balance of the evidence suggests that reasonably competitive conditions exist, notwithstanding the lack of variation in

finance rates. Furthermore, recent developments have reflected some tendency for credit card rates to decline.⁴⁹

A similar strand of thought is reflected in a recent *Wall Street Journal* editorial: "Credit-card interest almost certainly will come down. It will come down without rate ceilings. Nothing does it like competition."⁵⁰

Such arguments are insufficient. One cannot implicitly rely on the model of perfect competition as the principal defense for laissez-faire, given that the data cast severe doubt on the predictions of zero economic profits and cost-based pricing in this industry.

In order to make a cogent argument against regulation, one must proceed in a much more sophisticated fashion. First, it must be recognized that the behavior of the unregulated credit card market of the 1980's deviates in systematic ways from competitive predictions. The price of credit far exceeds its fundamental marginal cost, and the industry expects this situation to persist for some time. While nonprice competition has so far failed to impair firms' profits seriously, it appears to be steadily escalating, meaning that one can envision a day in the not-too-distant future when economic profits from new customers would be completely competed away via nonprice means. (Banks might still earn significant economic rents from their existing "captive" customers.)

Second, it should equally be recognized that regulation has only a limited potential to improve the outcome. The principal difficulty is that consumers occupy a spectrum of levels of credit-worthiness. Let ρ_n denote the bank's best estimate of the n th con-

⁴⁸ Possible reregulation of credit card interest rates has been the subject of controversy in recent years. In the 1987 Congressional session, no fewer than five bills dealing with credit cards were introduced: Senate Bill S.241 (mandating certain disclosures), S.242 (setting a national ceiling of four percent above the Internal Revenue Service's interest rate), S.616 (disclosure), S.674 (a ceiling of six percent above the Federal Reserve's discount rate), and House Bill H.R.515 (a ceiling eight percent above the one-year Treasury bill rate). In 1987-1988, Congress rejected all proposed bills and amendments setting credit card interest rate ceilings but enacted a mandatory disclosure bill.

⁴⁹ "Credit Card Interest Rates," Hearing before the Subcommittee on Consumer Affairs and Coinage of the Committee on Banking, Finance, and Urban Affairs, House of Representatives, Ninety-Ninth Congress, First Session, on H.R. 1197 and H.R. 3408 (October 29, 1985), Serial No. 99-44, U.S. Government Printing Office, Washington, 1986, page 39; also see *Federal Reserve Bulletin*, March 1986, p. 184.

⁵⁰ *Wall Street Journal*, March 16, 1987, editorial page.

sumer's probability of default and let c denote the marginal cost of lending funds to a consumer (exclusive of default risk). Then, the social optimum has every consumer paying his own, individualized interest rate: consumer n holds a credit card bearing a finance charge of $(\rho_n + c)$.⁵¹ Since conventional usury laws do not set interest rates according to the individual's default risk, ρ_n , they necessarily lead to outcomes that fall short of the optimum. The regulation is typically written: no bank is permitted to charge an interest rate greater than r^* . Under such a regulatory regime, no consumer whose default risk (to an external observer) exceeds $(r^* - c)$ will be serviced. Thus, if r^* is sufficiently low to ameliorate excess profits, it will also generally create deadweight loss by depriving individuals of the opportunity to hold credit cards. Moreover, it has been widely observed that, in an environment with price ceilings, there is a tendency for all firms to charge *exactly* the price ceiling. Hence, one could expect further deviation from the ideal, individualized interest rates: all consumers with default risks less than $(r^* - c)$ may end up paying interest rates equaling r^* .

A decision regarding the advisability of regulation thus involves a comparison of two less-than-ideal alternatives. The case for laissez-faire is strongest when one is only interested in efficiency and when non-interest-rate competition exclusively takes the form of recycling revenues to consumers. It has already been observed that credit card borrowers are highly interest-rate inelastic. Thus, high interest rates may not appreciably reduce the quantity borrowed, and so there may be little efficiency loss arising directly from excessive interest rates. The primary avenue for social loss is then the nonprice competition. However, to the extent that competition takes the form of frequent-flyer miles, cash rebates, or other relatively efficient means of recycling rev-

enues to consumers, there is still no appreciable social loss.

The case for regulation is strongest when one is upset by redistribution away from consumers or when nonprice competition expends substantial resources. I have already observed that high interest rates may be essentially neutral from an efficiency point of view. However, they presumably have a strongly undesirable redistributive effect from the comparatively poor (consumers who borrow on credit cards) to the comparatively rich (owners of bank stock). Moreover, there is a true (and potentially large) deadweight loss when nonprice competition takes the form of advertising.⁵² Some banks' reported noninterest expenses increased significantly from 1983 to 1988 even as the intrinsic cost of servicing accounts declined (e.g., Citibank, which advertises on national television); much of the additional expense probably represents marketing, and some fraction of this constitutes social loss.

IX. Conclusion

Despite the presence of 4,000 competitors, the bank credit card market of the 1980's behaved widely at variance with the predictions of a competitive model in continuous spot-market equilibrium. Interest rates approximated constancy, at levels around 18 percent per year, in the face of wide changes in banks' marginal costs. Profits persistently equaled three or more times the ordinary return on banking equity, with no sign of abatement. A breakdown of the optimizing consumer behavior so basic to the model of perfect competition may be an important element in the story.

The facts of the market are roughly consistent with a model of adverse selection in which many consumers are insensitive to

⁵¹ Either some banks could offer a spectrum of interest rates to different consumers, or banks could each offer just a single interest rate but specialize in consumers of different qualities of credit-worthiness.

⁵² The direct-mail credit card solicitations which I received at the rate of one per week while writing the article inspired this observation. One important, additional aspect of this problem is that the large interest-rate spread (see Section II-B) encourages banks to market cards in an aggressive way that makes them susceptible to fraud losses.

interest-rate differentials because they believe they will pay within the grace period (although they repeatedly fail to do so). This hypothesis is lent some empirical support by the finding that, assurances to the contrary, three-quarters of consumers pay finance charges on their outstanding credit card balances. Given the presence of such consumers, any bank that unilaterally reduced its credit card interest rate would disproportionately draw customers who actually do intend to borrow (i.e., the worst credit risks). Thus, the finance charges remain at high levels and become the main contributors to supranormal profits.

The facts of the market appear to be inconsistent with the predominance of well-informed consumers who are attempting to minimize their borrowing costs. There is no evidence that consumers are generally offered competitive interest rates on bank card balances, nor that most consumers respond to lower interest rates when they are offered.

The empirical findings of this article suggest a broader question: is it that the bank credit card market of the 1980's was uniquely pathological, or can one identify other markets whose structures seem equally conducive to the competitive model but whose empirical outcomes are similarly noncompetitive? This would seem to be a ripe area for further research.

APPENDIX A: THE BANK CREDIT CARD SURVEY

In May 1986, a pilot survey was mailed to 32 banks which were believed to be among the 50 largest bank issuers of credit cards. Five responses were received. The bank credit card survey (BCCS) was formed by using these responses to refine the questions asked. The BCCS was mailed in November 1986 to each of the 50 largest bank issuers of credit cards, as ranked in the *Nilson Report* (Number 371, January 1986), plus five banks ranked numbers 51-60. (The BCCS was not sent to the five banks that had responded to the pilot survey.) Following reminder letters in December 1986 and March 1987, as well as reminder telephone calls, 16 responses to the BCCS were received. Thus, the pilot survey and BCCS together elicited a total of 21 responses from a sample consisting of the following 58 banks:

Associates National Bank
Avco National Bank

BancOhio
Bank of America
Bank of New York
Bank One
Barnett Bank
Beneficial National Bank
Chase Manhattan Bank
Chemical Bank
Citibank
Citizens & Southern National Bank
Comerica Bank
Commerce Bank
Connecticut Bank and Trust Co.
CoreStates Bank of Delaware
Crocker National Bank
European American Bank
First City Bank
First Interstate Bank
First National Bank of Atlanta
First National Bank of Boston
First National Bank of Chicago
First National Bank of Omaha
First Omni Bank
First Tennessee Bank
First Wisconsin National Bank
Harris Trust and Savings Bank
Indiana National Bank
InterFirst Bank
Manufacturers Hanover Trust Co.
Marine Midland Bank
Maryland Bank
MBank
Mellon Bank
Mercantile Trust Co.
Michigan National Bank
National Bank of Detroit
National Westminster Bank
NCNB
Norwest Bank
PNC National Bank
Rainier National Bank
RepublicBank Dallas
Rocky Mountain Bankcard System
Seattle First National Bank
Security Bank and Trust Co.
Security Pacific National Bank
Signet Bank
Southeast Bank
Sovran Bank
State Street Bank and Trust Co.
Sun Bank
United Bank of Denver
United States National Bank
United Virginia Bank
Valley National Bank
Wells Fargo Bank

Several of the listed banks had ceased to exist as credit card issuers by the relevant time period, due to merger of the banks or acquisition of their portfolios. Seventeen responses included full interest rate series for the years 1982-1986 (see Table 2 for sizes), extended

TABLE A1—BANK CREDIT CARD SURVEY

1. Please indicate the interest rate, beginning in 1976 and through the present, on your most widely issued bank credit card:

Name of Card: _____

Year	February	May	August	November
1976	_____	_____	_____	_____
...
1986	_____	_____	_____	_____

Any additional information (for example, a different rate on premium cards, a floating-rate formula which you currently use, etc.):

2. Same as 1, for annual fee.
3. Please briefly describe the method your bank uses to compute bank card finance charges (include grace periods, etc.):
4. Please list charges other than annual fees (e.g., transaction charges, late fees, minimum finance charges, etc.) which your bank has charged between 1976 and the present. Please indicate relevant dollar amounts and dates:
5. Please list all major state and federal regulations (e.g., interest rate ceilings, laws prohibiting annual fees, etc.) which have hampered your operations between January 1976 and the present, indicating effective dates:
6. If your bank has any statement or position paper on credit card regulation, please enclose it with the completed survey.
7. Please indicate your number of total accounts, number of active accounts, total outstanding balances (at June 30 of each year, or another standardized date), annual charge volume, and charge-off rate:

Year	Number of total accounts	Number of active accounts	Outstanding balances	Annual charge volume	Charge-off rate
1976	_____	_____	_____	_____	_____
...
1986	_____	_____	_____	_____	_____

8. Please provide the following information about your cardholders, indicating for each column which of two possible pieces of information you are providing. [If both are available, please provide (A).]

Column 1:

- _____ (A) In an average month, what percent of your active accounts pay off their full outstanding balances (and so are not subject to a finance charge on those balances)?
- _____ (B) What percent of your active accounts pay off their full outstanding balances at least 11 months per year (and so are only subject to a finance charge on their balances at most one month per year)?

Column 2:

- _____ (A) Of your active accounts with outstanding balances, what is the average outstanding balance?
- _____ (B) Of all active accounts, what is the average outstanding balance?

Year	Percent who pay in full	Average outstanding balance
1976	_____	_____
...
1986	_____	_____

9. Please enclose a copy of the credit card application/solicitation(s) your bank uses most.
10. Please indicate which, if any, of the following factors you emphasize in the marketing of your cards.
- _____ Our high credit limit
 - _____ New customers can transfer their existing credit card balances onto our account
 - _____ New customers are waived our first year's annual fee
 - _____ Our interest rate is lower than our competitors'
 - _____ Our annual fee is lower than our competitors'
 - _____ Our card gives "bonus dollars" with each dollar charged, for discounts on merchandise
 - _____ Pre-approved credit card applications
 - _____ Free airline insurance
 - _____ Other freebies—list them:
 - _____ Other factors—list them:
11. Feel free to include any additional comments, either below or on separate sheets of paper.

TABLE A2—FOLLOW-UP BANK CREDIT CARD SURVEY (AND FOLLOW-UP SURVEY II)

1. 1987 (1988) updates of Questions 1 and 2 from Bank Credit Card Survey.
2. 1987 (1988) updates of Question 7 from Bank Credit Card Survey.
3. 1987 (1988) updates of Question 8 from Bank Credit Card Survey.
4. Please enter all available dollar figures for your bank's *credit card business only*. [Question 4 was patterned after the Federal Reserve System's *Functional Cost Analysis*. A xerox copy of p. 38 of the 1986 report was enclosed.]

	1984	...	1987 (1988)
1. Average total outstanding balances:	\$ _____	...	\$ _____
Income:			
2. Finance charge interest and customer fees (including annual fee):	\$ _____	...	\$ _____
3. Merchant discount, interchange fees, and other income:	\$ _____	...	\$ _____
4. Total income (2 + 3):	\$ _____	...	\$ _____
Operating expenses:			
5. Marketing and advertising:	\$ _____	...	\$ _____
6. Enhancements and affinity program expenses:	\$ _____	...	\$ _____
7. All other expenses (including salaries, fringe benefits, data services, processing, franchise fees; excluding items below):	\$ _____	...	\$ _____
8. Total operating expenses (5 + 6 + 7):	\$ _____	...	\$ _____
Earnings:			
9. Net earnings before losses (4 - 8):	\$ _____	...	\$ _____
10. Net credit losses:	\$ _____	...	\$ _____
11. Net fraud losses:	\$ _____	...	\$ _____
12. Net earnings (9 - 10 - 11):	\$ _____	...	\$ _____
Memoranda:			
13. Cost of funds:	\$ _____	...	\$ _____
14. Net earnings (pretax) after cost of funds (12 - 13):	\$ _____	...	\$ _____

through 1987 by contemporaneous telephone calls. Respondents were promised anonymity.

The follow-up bank credit card survey was mailed to the 21 initial respondents in January 1988, requesting both 1987 updates of data that the original survey had elicited and direct reports of credit card profits. Follow-up survey II was mailed to the 21 respondents in February and July 1989, requesting both 1988 updates of data that the original survey had elicited and direct reports of credit card profits. Following reminder letters, 11 responses were received, seven of which contained data on profits for 1984–1987 (see Table 2 for sizes) and five also for 1988. The profit reports of Bank F and all 1988 profit reports were completed by banks after the working-paper precursor of this article was made available to the banks. Respondents were again promised anonymity. The BCCS and follow-up BCCS are reprinted in Tables A1 and A2 in condensed form.

APPENDIX B: PROFITABILITY CALCULATIONS

Bank Credit Card Survey Data

COST OF FUNDS is defined by taking the one-year Treasury-bill yield plus 0.75 percent, averaged over the calendar year, and multiplying by 0.94. The Treasury-bill yield is taken from the *Federal Reserve Bulletin*, April 1990, table 1.35, line 21 (and previous issues). The number 0.75 represents the spread between yields on Treasury securities and yields on credit-card-backed securities. The number 0.94 represents 1 minus the banking system's capital requirement of 6 percent.

The numbers reported in Table 6 were constructed as follows (see also Table A2):

$$\text{ROA (reported)} = \text{BCCS Line 14} / \text{BCCS Line 1}$$

$$\begin{aligned} \text{ROA (adjusted)} = & (\text{BCCS Line 14} + \text{BCCS Line 13} \\ & - \text{COST OF FUNDS} \times \text{BCCS} \\ & \text{Line 1}) / \text{BCCS Line 1} \end{aligned}$$

$$\text{ROE (adjusted)} = \text{ROA (adjusted)} / 0.06.$$

Call Report Data for Credit Card Banks

The calculations reported in Table 7 are based on the quarterly consolidated reports of condition and income ("call reports") which "credit card banks" filed with the FDIC. Included in the sample were all commercial banks that met both of the following criteria:

- 1) credit card balances constituted at least 75 percent of the bank's total assets (so that the bank's profits are a good proxy for profits attributable to the credit card business);
- 2) the bank's balance sheet was not seriously marred by credit card securitizations or portfolio acquisitions.

For example, Maryland Bank was excluded from the sample in 1988, because that bank's credit card balances averaged \$3.1 billion in that year, while only \$1.7 billion appeared on the bank's Report of Condition (the remainder having been securitized). Typically, for banks in Table 7, credit card balances constituted 97 percent or more of total assets.

In the description immediately below, the December 31, 1987 call report for Citibank (South Dakota), N.A., is used to standardize line numbers. The following data were extracted from credit card banks' call reports:

AVERAGE TOTAL ASSETS = Schedule RC-K, line 9 [Total Assets] – Schedule RC, line 10 [Intangible Assets] (this calculation is performed for each of the March 31, June 30, September 30, and December 31 reports; I work with the arithmetic average of the four numbers);

TOTAL EQUITY PREVIOUS YEAR = Schedule RI-A, line 3 [Amended Balance End of Previous Calendar Year];

INCOME BEFORE TAXES = Schedule RI, line 8 [Income Before Income Taxes and Extraordinary Items];

PROVISION FOR LOAN LOSS = Schedule RI, line 4A [Provision for Loan and Lease Losses];

NET CHARGEOFFS = Schedule RI-B, line 9, column A [Total Charge-Offs] – Schedule RI-B, line 9, column B [Total Recoveries];

INTEREST EXPENSE = Schedule RI, line 2F [Total Interest Expense];

COST OF FUNDS = as in first paragraph of this appendix;

AMORTIZATION EXPENSE OF PREMIA = Schedule RI-E, line 2A [Amortization Expense of Intangible Assets].

The numbers reported in Table 7 were then constructed as follows:

ROA (reported) = (INCOME BEFORE TAXES + PROVISION FOR LOAN LOSS – NET CHARGEOFFS) / AVERAGE TOTAL ASSETS;

ROE (actual cap) = (INCOME BEFORE TAXES + PROVISION FOR LOAN LOSS – NET CHARGEOFFS) / TOTAL EQUITY PREVIOUS YEAR;

ROA (adjusted) = (INCOME BEFORE TAXES + PROVISION FOR LOAN LOSS – NET CHARGEOFFS + INTEREST EXPENSE – COST OF FUNDS × AVERAGE TOTAL ASSETS + AMORTIZATION EXPENSE OF PREMIA) / AVERAGE TOTAL ASSETS;

ROE (adjusted) = ROA (adjusted) / 0.06.

Prospectus Data

The calculations reported in Section III-D are based on information contained in prospectuses and registration statements filed with the SEC in connection with all public credit card securitizations by commercial banks from 1987 to early 1990. The following is a complete list of the banks and the prospectuses used.

- 1) BancOhio (National City Corporation): National City Credit Card Trust 1990-A, registration statement dated January 2, 1990, pp. 17, 23;
- 2) Bank of America: California Credit Card Trust 1987-A, prospectus dated February 25, 1987, pp. 10, 14; California Credit Card Trust 1987-B, prospectus dated June 19, 1987, pp. 11, 14;
- 3) Chemical Bank: Chemical Bank Credit Card Trust 1988-A, prospectus dated August 16, 1988, pp. 14, 15; Chemical Bank Credit Card Trust 1989-A, prospectus dated October 30, 1989, pp. 16, 19;
- 4) Citibank (South Dakota)/Citibank (Nevada): Money Market Credit Card Trust 1989-1, prospectus dated January 25, 1990, pp. 21, 23;
- 5) Colonial National Bank U.S.A.: Colonial Credit Card Trust 1988-A, preliminary prospectus dated March 23, 1988, pp. 18, 23;
- 6) First National Bank of Chicago/FCC National Bank: First Chicago CARDS Trust 1987-1, prospectus dated September 29, 1987, pp. 13, 15; First Chicago Master Trust, registration statement dated November 16, 1989, pp. 16, 17;
- 7) Lomas Bank U.S.A.: Lomas Credit Card Trust 1989-A, registration statement dated July 3, 1989, pp. 17, 24;
- 8) Manufacturers Hanover Trust Company: MHARCCS Trust 1988-1, prospectus dated June 21, 1988, pp. 17, 18 (1988 only through March 31);
- 9) Maryland Bank, N.A.: MBNA Credit Card Trust 1988-B, prospectus dated September 9, 1988, pp. 16, 21; MBNA Credit Card Trust 1989-B, registration statement dated November 8, 1989, pp. 16, 20;
- 10) RepublicBank Delaware: securitization of January 16, 1987, as summarized in *Standard & Poor's Asset-Backed Securitization CreditReview*, March 16, 1987, pp. 21, 22;
- 11) Southeast Bank: Southeast Bank Credit Card Trust 1990-A, registration statement dated January 29, 1990, pp. 17, 22.

Premia Paid for Credit Card Portfolios

Table 9, the list of premia paid for credit card portfolios, reflects manual and computerized searches of national newspaper indexes over the period January 1984–April 1990. In order to be included, a transaction was required to meet all of the following criteria:

- 1) the transaction was reported in the *Wall Street Journal*, *The New York Times*, *The American Banker*, or the Kidder, Peabody & Co. report (Kristina E. Andersson and Alison A. Deans, 1989);
- 2) the exact premium, the parties to the transaction, the approximate date of the transaction, and the approximate size of the portfolio were reported;
- 3) the transaction was essentially an unbundled sale of credit card accounts and nothing else.

In the event of conflicting reports, the conflict was resolved using the best available information.

REFERENCES

- Andersson, Kristina E. and Deans, Alison A., *Credit Cards: How to Pick a Winner in a Consolidating Industry*, industry report, New York: Kidder, Peabody & Co., September 1989.
- Arak, Marcelle, Englander, A. Stephen and Tang, Eric M. P., "Credit Cycles and the Pricing of the Prime Rate," *Federal Reserve Bank of New York Quarterly Review*, Summer 1983, 8, 12–18.
- Ausubel, Lawrence M., "Oligopoly When Market Share Matters," mimeo, Stanford University, May 1984.
- , "The Failure of Competition in the Credit Card Market," Banking Research Center Working Paper No. 153, Northwestern University, October 1988.
- Berlin, Mitchell and Mester, Loretta J., "Credit Card Rates and Consumer Search," mimeo, Federal Reserve Bank of Philadelphia, February 1988.
- Canner, Glenn B. and Fergus, James T., "The Economic Effects of Proposed Ceilings on Credit Card Interest Rates," *Federal Reserve Bulletin*, January 1987, 73, 1–13.
- Diamond, Peter, "A Model of Price Adjustment," *Journal of Economic Theory*, 1971, 3, 156–68.
- Douglas, George W. and Miller, James C., *Economic Regulation of Domestic Air Transport: Theory and Policy*, Washington, DC: Brookings Institution, 1974.
- Farrell, Joseph and Shapiro, Carl, "Dynamic Competition with Lock-In," Department of Economics Working Paper No. 8727, University of California, Berkeley, January 1987.
- Klemperer, Paul, "Markets with Consumer Switching Costs," *Quarterly Journal of Economics*, May 1987, 102, 375–94.
- Pavel, Christine and Binkley, Paula, "Cost and Competition in Bank Credit Cards," *Economic Perspectives* (Federal Reserve Bank of Chicago), March/April 1987, 11, 3–13.
- Rotemberg, Julio J. and Saloner, Garth, "The Relative Rigidity of Monopoly Pricing," *American Economic Review*, December 1987, 77, 917–26.
- Stiglitz, Joseph E. and Weiss, Andrew, "Credit

- Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, 71, 393-410.
- White, James J. and Munger, Frank W., "Consumer Sensitivity to Interest Rates: An Empirical Study of New-Car Buyers and Auto Loans," *Michigan Law Review*, June 1971, 69, 1207-58.
- Asset-Backed Securitization CreditReview, supplement to *CreditWeek*, New York: Standard & Poor's, March 16, 1987.
- Consolidated Reports of Condition and Income, Washington, DC: Federal Deposit Insurance Corporation, quarterly, 1983-1988.
- Credit Card Interest Rates, House of Representatives Subcommittee Hearing, Washington, DC: U.S. Government Printing Office, Serial No. 99-44, 1986.
- Credit Card Management, New York: Faulkner & Gray, monthly, 1988-1989.
- Credit Card News, New York: Faulkner & Gray, biweekly, 1988-1989.
- Federal Reserve Bulletin, Washington, DC: Board of Governors of the Federal Reserve System, monthly, 1982-1990.
- Functional Cost Analysis, National Average Reports, Commercial Banks, Washington, DC: Board of Governors of the Federal Reserve System, annual, 1976-1988.
- Nilson Report, Los Angeles: HSN Consultants, biweekly, 1986-1989.
- Prospectuses and Registration Statements, Washington, DC: Securities and Exchange Commission, irregular, 1987-1990.

Externalities and Growth Accounting

By JESS BENHABIB AND BOYAN JOVANOVIĆ*

This paper tackles two puzzles: the high empirical elasticity of aggregate output with respect to the measured capital input and the seemingly high variability of growth rates over countries in the medium run. We find that one need not invoke increasing returns or externalities to capital to explain these two puzzles. Rather, they are consistent with a constant-returns-to-scale aggregate production function, so long as the exogenous Solow residual process has enough persistence in it. In our model, causality runs exclusively from knowledge to capital, and therefore the apparent absence of an external effect to the capital input says nothing about the importance of spillovers in the creation of knowledge. (JEL 110)

This paper addresses the question of how to explain the variation in the levels and rates of growth of output across countries. It focuses in particular on the question of whether or not this cross-country variation offers support for the suggestion that there are aggregate increasing returns to capital and labor caused either by external effects associated with capital investment or by a secular increase in the variety of intermediate inputs. By looking again at the evidence considered by Paul Romer (1987) and Lawrence Christiano (1987), we show that, under plausible assumptions about the behavior of the economy, there is no support for the assertion that capital-related externalities are present. We show instead that the variation in countries' growth rates is consistent with each country having the same constant-returns-to-scale production function and with a stochastic process for technological change that is the same across different countries but starts from different initial positions.

A. The Issues

Two issues are at hand. The first concerns the prediction of Robert Solow's (1957) model that the elasticity of output with respect to capital should equal capital's share in output, which is roughly one-third. Yet when one looks at data over longer periods of time and in several different countries, as Romer (1987) has, one finds an elasticity that is closer to unity. There are three extensions of Solow's model that can generate this. One is to make the identifying assumption that Romer makes, at least implicitly, that the fundamental exogenous variation is in the rate of savings and investment. In this case, one must change the elasticity of output with respect to capital by invoking capital-related external effects or increasing returns stemming from input variety. The second possibility is to make the identifying assumption that Christiano (1987) makes and that we discuss in Section IV, namely that the exogenous variation across countries is in the underlying exogenous process of technical change. The third alternative, suggested here, is to assume that the fundamental stochastic process for technology is the same but that in the sample the realized paths differ across countries. In this case, something like Christiano's explanation can be constructed (and no capital externalities are needed), but this can be done without invoking his unattractive "fixed effects" as-

*Professors of Economics, New York University, 269 Mercer Street, New York, NY 10003. We thank The C. V. Starr Center for Applied Economics for financial and technical assistance, Francesco Goletti and Ray Atje for research assistance, and Bob Barro, Will Baumol, Larry Christiano, Stan Fischer, Zvi Griliches, Ned Nadiri, Ariel Pakes, and Paul Romer for helpful comments. An earlier version was given at an NBER growth conference in Cambridge, Massachusetts, in October 1989.

sumption that there are permanent, exogenous differences in the technology faced by different countries. Rather, the difference across countries lies in the initial conditions and implicitly in the sequence of historical accidents that lead to those conditions. We deal with this first issue in Section IV.

The second issue is whether (our version of) Solow's model can be reconciled with the seemingly large variation in countries' growth rates since World War II. We shall argue that this variation is roughly consistent with Solow's model. Part of the new evidence that we bring to bear on this assertion is the finding that, for the population of countries as a whole, the time-series variation in output growth in a representative country is consistent with the cross-country variation in output growth measured over 25 years. This bears on the question of the inherent differences, if any, that must be invoked to explain the disparate behavior of the different countries. We deal with this second issue in Section III.

B. Capital and Knowledge

There seems, on first consideration, to be little reason to expect a firm's investment in capital to have substantial beneficial spillover effects in reducing production costs of other firms. However, if firms with more capital also have more productive knowledge and if this knowledge spreads to other firms, then unless one can somehow measure knowledge and control for it, an increase in the capital stock of one firm will appear to lower the production costs of other firms. In the same vein, if the economy's capital stock is positively related to the availability of specialized intermediate inputs and if these inputs are not measured, the growth of capital will appear to increase aggregate output by more than its private marginal product.

In his pioneering article, Romer (1987) offers two separate models each of which can generate a capital elasticity of output larger than one-third. Within the context of his first model, he argues that a large positive externality in capital formation is

needed to explain the strong positive association in aggregate data (over countries and over epochs) between the "Solow residual" and the growth of the capital stock. Moreover, the size of his externality estimate is staggering: the social marginal product of capital, suggests Romer, is perhaps twice or even three times its private marginal product, and by implication, the equilibrium level of investment falls far short of its socially optimal level.¹

Romer's second model introduces knowledge explicitly in the form of ideas for intermediate goods. In its implications for comovements between aggregate output, capital, and labor, this model is observationally equivalent to a version of his first model, as is evident in his equation 11. This model need not have any *direct* spillovers of knowledge to yield aggregate increasing returns, although in later versions Romer introduces direct spillovers in the research sector. The "externality" that the final goods sector enjoys is, as Romer points out, a pecuniary one, and a divergence between equilibrium and social optimum could arise solely because of the monopoly power introduced into the intermediate-goods sector so as to provide incentives for inventions.

Romer's two models have a common feature: growth in capital *causes* a growth in knowledge or a growth in the availability of specialized inputs, or both.² Evidence of direct spillovers of knowledge then constitutes support for Romer's first model. On the other hand, evidence that *pecuniary* spillovers are present (and that such spillovers increase with the capital stock) would provide support for the second model, so long as some increasing returns (as inputs *and* variety vary) are also present. Un-

¹Christiano (1987) challenges these conclusions, claiming that a balanced-path outcome for the Solow model is consistent with the data and that no capital externality is required. Indeed, in the deterministic case, along the balanced-growth path, the externality cannot be identified. Martin N. Baily (1987) and others have made the same point. We return to Christiano's argument in Section IV.

²See especially Romer's reference to evidence from Jacob Schmookler (1966) to the effect that in various industries patenting tends to *follow* investment.

fortunately, the available evidence often mixes the two types of externality (pecuniary and nonpecuniary) so that their relative (as well as absolute, it turns out) importance is hard to pin down.

We begin with the assumption, implicit in Romer's first model, that an increase in a firm's capital stock causes the firm's productive knowledge to go up in the same proportion, so that we can use estimates from micro data on externalities in R&D as an estimate of the size of the capital externality. If direct spillovers to R&D do exist, they are likely to be largest among firms in the same industry, since those firms are likely to be using similar technologies. The largest *intraindustry* spillover estimates were obtained by Jeffrey Bernstein and Ishaq M. Nadiri (1989), who find that, in four industries, the social returns to intraindustry spillovers of R&D ranged from 30 percent to 123 percent of the private returns to R&D.³ Such large estimates are, however, exceptions: in a summary of the literature

on the elasticity of output with respect to the R&D input, Zvi Griliches (1988 p. 15) reports that "while the presence of spillovers would make one expect the industry-level coefficients to be higher than those estimated at the firm level, the econometric estimates do not show this in any convincing fashion."⁴ If aggregating up to the industry level makes little difference to the estimates of the R&D coefficient, it would be quite surprising if aggregating to the whole economy would produce a large upward revision (specifically, a tripling) of the R&D coefficient. However, this is exactly what Romer's first argument implies, and the micro data do not seem to support it.

The micro evidence does not seem to favor Romer's second argument either. Frederick M. Scherer (1982) finds that productivity growth in an industry is strongly correlated with the extent to which it purchases R&D-embodied products. Additionally, a wealth of evidence points to sustained cost reductions in a whole range of intermediate-inputs services. It is beyond dispute, therefore, that final-goods producers have benefitted from sustained improve-

³Edwin Mansfield et al. (1977) also find large spillovers for a select group of innovations, but since these were all *successful* innovations, their sample does not accurately represent the outcome of investments in R&D. On the other hand, while the absolute value of the private and social rates of return is clearly biased upward in their sample, their *relative* magnitudes are perhaps not biased. Among their 18 innovations, the social rate of return averaged 77 percent, while that of the private rate averaged 33 percent. These results do support Romer's claim that the social returns might exceed the private rate by a factor of more than two. Bernstein and Nadiri's results must also be viewed with caution, because they are based on a deterministic model, and simultaneity biases are likely to be present because of omitted time effects and unobservable industry effects. In essence, they evaluate the spillover from the partial correlation between a firm's investment in physical capital and R&D on the one hand and the industry investment in R&D on the other. These variables will usually be positively correlated, because they both will respond to industry shocks, time effects, and so on, and this response will cause an upward bias on any estimate of R&D spillovers that relies on this partial correlation. Similarly, Adam Jaffe's (1986) finding that there were significant spillovers in a cross section is questionable on grounds that his assumption (on p. 992) about an absence of correlation between his instrumental variables and the unobserved "technological opportunity" parameter that each firm faces is unlikely to be met.

⁴Ariel Pakes and Mark Schankerman (1984b) find a much stronger correlation between industry-wide R&D and lagged industry growth than they do between firm R&D and firm growth. However, they interpret the causality as running from industry growth to R&D, in the spirit of Schmookler's argument that the incentive to do R&D increases as market size grows. It seems crucial, at the industry level, to impose assumptions that allow one to distinguish shifts in product-demand from shifts in technological opportunity. A different, and more questionable, source of evidence on spillovers is the rate at which the economic value of private knowledge decays. The faster a piece of knowledge spills over to other firms, the faster, presumably, is the loss of economic rent that the firm can extract from that piece of knowledge. Pakes and Schankerman (1984a) find that knowledge depreciates much faster than physical capital, although they do not interpret this as implying a high spillover rate for knowledge. Unfortunately, as Griliches (1979) points out, the value of private knowledge may decay not just because it "leaks" to other firms but also because it is superseded by new knowledge generated by other firms. In other words, the economic value of knowledge would depreciate even in a world with no spillovers, and its depreciation rate is thus an unreliable indicator of the extent and speed of spillovers.

ments in input quality. This does not mean however, that there are increasing returns in the aggregate. In fact, we know of no evidence that (a) aggregate returns to measured inputs *and* variety increase or (b) the provision of variety is fueled by a larger stock of capital.

In short, the micro data are so far not conclusive on Romer's hypotheses. Our aim here is to show that *aggregate* data are consistent with a view that neither direct nor pecuniary spillovers are fueled by physical capital. In doing this, we leave open the possibility that there are increasing returns due to something else—human capital, perhaps.

C. Our Argument

In our model, causality runs entirely from knowledge to capital. Knowledge evolves exogenously; we do not estimate its external effects, and indeed, under our assumption about causality, micro evidence in spillovers of knowledge says *nothing* about spillovers to the capital input. The popular view that some capital investment is needed for the implementation of new ideas favors our causality assumption, since it is natural (as in Andrei Shleifer [1986], for instance) to imagine that new ideas precede the installation of the capital equipment needed to implement them. Moreover, at the level of the individual firm at least, micro data indicate that R&D Granger-causes investment, but that investment does not Granger-cause R&D (Saul Lach and Mark Schankerman, 1989).

While it reverses the assumption about causality between capital and knowledge, our model still admits the possibility of an externality to the capital input and is in fact almost the same as Romer's. Our conclusions, however, are quite different: we examine a variety of bodies of data and find no evidence to support the hypothesis that there are beneficial spillovers arising from the capital input. The reason why our conclusions differ from Romer's is roughly as follows. Romer faces simultaneity problems when he estimates a production function in which capital and labor are endogenous and

correlated with the disturbance to the production function. One source of disturbances to the production function is the business cycle, and Romer tries to remove it by filtering out the high frequencies with long-run averages. He further recognizes that, even in the long-run data, low-frequency movements in technology might create a correlation between the inputs and the production-function disturbance, but he argues⁵ that the extent of this correlation could not plausibly be so large as to reverse his conclusions. This, however, is where we disagree with his argument. We make explicit assumptions about the way in which the capital and labor inputs evolve in response to changes in the state of technology. These assumptions enable us to calculate the correlation between the inputs and the disturbance. We find that, even in the long-run data, this correlation is plausibly high enough to explain the high empirical elasticity of output with respect to the capital input. Moreover, the positive association between knowledge shocks and capital investments also seems to explain most of the variance in countries' growth rates in the postwar period. No externalities or increasing returns are needed.

Section I presents our model, which consists of five structural equations. In Section II, we present maximum-likelihood and least-squares estimates for the model using postwar quarterly and annual U.S. data and find no evidence of an externality. In Section III, we then discuss some of the model's implications about the convergence of GNP among different countries, and interpret the apparent empirical validity of "Gibrat's Law" in the behavior of countries' GNP series over extended periods. In Section IV, we reinterpret Romer's regression results (which use data on growth of inputs and output over long epochs) in terms of the simultaneity biases that we calculate, and we conclude that even those data offer no

⁵Especially on p. 194 with reference to evidence on the persistence of cross-country differentials in growth rates.

evidence for the conjectured positive externality to the capital input.

After the empirical evidence discussed in Sections II–IV, Section V presents two models that give rise to the structural equations first introduced in Section I. The first is a stochastic Diamond type of overlapping-generations model, the second a Brock-Mirman type of infinite-horizon model. The sixth and final section offers some concluding remarks.

I. The Augmented Solow Model

The representative firm produces output Y_t with hired inputs K_t and L_t , taking as given the economy-wide capital stock \bar{K}_t per firm and the state of knowledge Z_t . The production function is

$$(1) \quad Y_t = K_t^\alpha L_t^{1-\alpha} \bar{K}_t^\theta Z_t.$$

In the first version of Romer's model, the parameter θ measures the external effect of capital, an effect that the firm ignores when making its decisions. In the second version, θ represents increasing returns in the variety of intermediate inputs whose quality is (in an auxiliary equation) linked to the economy-wide capital stock. Since all firms are the same, $K_t = \bar{K}_t$. Letting lowercase symbols denote logarithms, (1) reads

$$(2) \quad y_t = (\alpha + \theta)k_t + (1 - \alpha)l_t + z_t.$$

When $\alpha + \theta$ is unity, this equation is the same as Romer's equation 11. If the firm is a price-taker in its product and factor markets, α is capital's share in output, and $1 - \alpha$ is labor's share. This is Romer's reformulation of Solow's model.⁶

To this, we now add assumptions about how knowledge grows and about how the equilibrium k_t and l_t evolve. Knowledge evolves exogenously, as follows:

$$(3) \quad z_{t+1} = \mu + \rho z_t + \omega_t \quad |\rho| \leq 1$$

$$(4) \quad \omega_t = \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2}.$$

⁶At least a part of this model is in Griliches (1979 p. 102); Griliches there attributes it to an unpublished note by Yehuda Grunfeld and David Levhari.

Thus, the z_t process is an ARMA(1,2). Three features of (3) and (4) deserve mention. First, μ is the rate of exogenous technical change; one expects it to be positive, since some knowledge comes for free from abroad, and in addition, some knowledge is generated for free domestically as a by-product of everyday economic activity. At any rate, the economy's long-run growth rate in GNP per capita will be $\mu(1 - \alpha)$ if ρ is unity and 0 if ρ is less than 1. On these grounds, we expect ρ to be roughly equal to unity. Second, the parameter ρ , or rather $1 - \rho$, measures the rate at which knowledge depreciates.⁷ From the work on business cycles by Edward C. Prescott and his coworkers (e.g., Prescott, 1986), we expect ρ to be about 0.99 in the quarterly U.S. data and 0.96 in the annual data. An important omission here is a purely transitory component to z . Its inclusion would almost certainly raise the estimates of ρ that we present in the next section and would work in our favor. In other words, the inclusion of transitory monetary and other policy effects on the measured Solow residual and even of errors in measuring z_t would, as will become clear below, help rather than hurt our case, which hinges on ρ being close to 1. Third, the MA(2) specification for ω_t was entirely arbitrary; it took two moving-average terms to remove the autocorrelation from the residuals in the quarterly U.S. data.

Next we specify the behavior of the capital and labor inputs. In Section VI, we shall present two separate micro models⁸ that imply the following equation as governing competitive equilibrium allocations:

$$(5) \quad k_{t+1} = \gamma + y_t$$

where γ is a constant.⁹ For some of the

⁷Because it is superseded by other knowledge. Actually, it is not knowledge but rather its economic value that depreciates.

⁸One of these is in many respects similar to the model that Edward Prescott (1986) proposes for business-cycle analysis.

⁹Equation (5) is an exact specification for the infinite-horizon representative-agent model with logarithmic utility, Cobb-Douglas production and 100-percent depreciation of capital. The details are in Section V.

results in Sections II and IV, we also assume that

- (6) l_t is a stochastic process
independent of z_t .

This assumption, which is also given a micro justification in Section VI, is not used in Section III, however, where we interpret long-run growth differentials in countries' growth experience.

Equations (2)–(5) and assumption (6) make up the model. In sum, it is Romer's model with the added assumptions of exogenous knowledge and endogenous capital and labor. The next three sections describe how we have estimated its parameters. Section II uses postwar U.S. data, Section III uses Alan Heston and Robert Summers's (1984) data, and Section IV uses the longer-run data that Romer compiled from Angus Maddison (1982) and elsewhere. None of these bodies of data supports the hypothesis that θ is positive.

II. Estimates From Postwar U.S. Data

We begin our empirical inquiry by looking at the postwar U.S. data. The reader will not be surprised to learn these data offer no support for a positive θ , since (a) Romer himself did not cite these data as supportive of capital externalities, (b) Prescott (1986) has, with some success, used a model quite similar to ours but with θ set equal to 0 to fit detrended postwar U.S. data, and (c) the short-run data are notorious in that output fluctuations are explained almost entirely by variations in

hours worked and hardly at all by the measured capital input.

A problem presented by the capital input is that it is likely to be poorly measured, at least at high frequencies, because of variation in its utilization rate. Our estimation procedure in this section begins by treating k_t as unobservable. This assumption underlies the calculation of the estimates in the first four tables. Tables 5–8, on the other hand, do use capital data.

Substitution of (5) into (2) yields

$$(7) \quad y_t = (\alpha + \theta)\gamma + (\alpha + \theta)y_{t-1} \\ + (1 - \alpha)l_t + z_t.$$

This is the equation that formed the basis for the estimates reported in the first two tables, which used data on $\log(\text{GNP})$ for y_t and $\log(\text{hours worked})$ for l_t . The data are *not* detrended. We present two sets of estimates. Table 1 reports the unconstrained maximum-likelihood estimates¹⁰ for the annual and the quarterly data separately. Table 2 reports estimates for the remaining parameters when α is constrained to equal $\frac{1}{3}$ (i.e., capital's postwar share in income).

Several points are noteworthy. For the quarterly data, the unconstrained estimates are virtually the same as the constrained estimates, and the likelihood ratio does not significantly differ from one. The estimate of ρ is about the same as that of Prescott (1986 p. 15). The estimate of θ is close to 0 and does not differ significantly from 0. When α is freed up, its estimate does not

¹⁰The likelihood was derived as follows. Multiplying (7) through by ρ , lagging one period and subtracting the result from (7) yields

$$y_t - \rho y_{t-1} = C + (\alpha + \theta)(y_{t-1} - \rho y_{t-2}) \\ + (1 - \alpha)(l_t - \rho l_{t-1}) + \omega_t$$

where $C = (1 - \rho)[(\alpha + \theta)\gamma + \mu]$. The ε_t are assumed to be normally distributed. Since ω contains two moving-average terms, we used the Box-Jenkins procedure, setting the two presample error terms to their zero means. For y , we use $\log(\text{GNP})$ in 1982 dollars; for l , we use \log of total hours worked; and for k , we use the \log of total (private fixed plus government) capital, all for the period 1947–1985, or 1986.

However, for a model with less than 100-percent depreciation and general functional forms, the qualitative features of this relationship, that is the positive covariance of k_{t+1} with k_t as well as z_t , which are the critical elements driving our results in the following section, will be preserved under very reasonable assumptions. This issue is explicitly discussed in Section V (especially see Lemma 1 and the surrounding discussion). Moreover, in Section II we shall also present the estimates for the model's parameters when the evolution of the capital stock obeys $K_{t+1} = sY_t + (1 - \delta)K_t$ instead of (5) (see Tables 3 and 4). We shall also present estimates in Tables 5–8 that use capital data and hence bypass (5) and (5').

TABLE 1—UNCONSTRAINED MAXIMUM-LIKELIHOOD ESTIMATES, BASED ON (5)

Data	Parameter	Estimate	SE	<i>t</i>	<i>P</i>
Yearly ^a	ρ	0.92	0.02	38.84	0.000
	θ	0.01	0.09	0.14	0.887
	<i>C</i>	-0.36	0.09	-3.72	0.000
	λ_1	0.32	0.09	3.24	0.002
	λ_2	0.17	0.09	1.80	0.080
	α	-0.12	0.09	-1.23	0.225
	σ_e^2	0.00014			
Quarterly ^b	ρ	0.98	0.002	397.21	0.000
	θ	-0.13	0.034	-3.87	0.000
	<i>C</i>	-0.01	0.004	-3.72	0.000
	λ_1	-0.15	0.039	-3.99	0.000
	λ_2	0.19	0.042	4.51	0.000
	α	0.35	0.042	8.24	0.000
	σ_e^2	0.00007			

Note: The estimates are unconstrained in that α need not equal $\frac{1}{3}$.

^aLog likelihood = 101.63 (37 observations, 31 degrees of freedom).

^bLog likelihood = 547.86 (166 observations, 160 degrees of freedom).

TABLE 2—CONSTRAINED MAXIMUM-LIKELIHOOD ESTIMATES, BASED ON (5)

Data	Parameter	Estimate	SE	<i>t</i>	<i>P</i>
Yearly ^a	ρ	0.94	0.02	43.34	0.000
	θ	-0.53	0.14	-3.66	0.000
	<i>C</i>	0.10	0.08	1.23	0.227
	λ_1	0.61	0.30	2.03	0.050
	λ_2	0.38	0.26	1.45	0.155
	σ_e^2	0.00023			
Quarterly ^b	ρ	0.98	0.01	51.45	0.000
	θ	-0.11	0.04	-2.73	0.006
	<i>C</i>	-0.02	0.03	-0.54	0.585
	λ_1	-0.16	0.04	-3.81	0.000
	λ_2	-0.18	0.04	4.39	0.000
	σ_e^2	0.00007			

Note: The estimates we constrained in that $\alpha = \frac{1}{3}$.

^aLog likelihood = 94.09 (37 observations, 32 degrees of freedom).

^bLog likelihood = 547.82 (166 observations, 161 degrees of freedom).

differ significantly from $\frac{1}{3}$. All in all, then, the model does pretty well with the quarterly data.

Such is not the case with the annual data. The unconstrained estimate of α is negative, but not significant; and, θ is positive, but small and nonsignificant. Thus, the social marginal product of capital appears to be low in these data. This is especially clear in the first panel of Table 2 in which, when α is constrained to $\frac{1}{3}$, θ is large but negative and highly significant. The likelihood-ratio

test resoundingly rejects the restriction that $\alpha \equiv \frac{1}{3}$. These results with annual data are quite similar to the regression results that Romer reports in line 2 of his table 2; in this regression, he allows (as we do) for exogenous technical change and measures (as we do) the labor input by hours worked.

An important source of downward bias on θ deserves mention, however. While our specification (4) does allow for transitory components in y , there may nevertheless be measurement error in y that will cause the

coefficient of lagged y in (7), namely $\alpha + \theta$, to be underestimated because of errors-in-variables bias. In Table 2, this will cause θ to be underestimated, while in Table 1, where α is freed, both α and θ may be underestimated. Put differently, measurement error in y will lead to a spurious negative dependence between the quasi-first differences in footnote 10. Such a negative bias could hide a positive θ .

A second set of problems arises because it takes time to build capital, and it also takes time for the external benefits of capital accumulation to be felt. That is, it is possible not only that there are significant building-time delays, but that externalities affect output with a lag. To test for the presence of such delays, we considered a production function $Y_t = K_{t-p}^\alpha L_t^{1-\alpha} K_{t-s} Z_t$, where p and s represent lags (presumably, $0 \leq p \leq s$). We derived the corresponding reduced form [the analogue of (7)] for the infinite-horizon representative-agent model where y on the right-hand side appeared with lags s and p . We estimated this model using quarterly data for various values of s and p and found that the best fits, in terms of likelihood, were for $s = p = 0$. For all values that we checked for $s > 0$ (up to $s = 12$) and $p = 0$, the externality coefficient θ was 0, and for values $s \equiv p > 0$, it tended to be negative.

A third set of problems surround our assumption in equation (5) about the way in which capital evolves. Distinct from the issues discussed above concerning the length of time that it takes to build capital, there is the issue of how long capital remains productive after it has been built; that is, how fast it depreciates. If y_t is measured by GNP, as we have done, then equation (5) implies that there is 100-percent depreciation. On the other hand, if Y_t is interpreted as wealth, then we have not measured wealth correctly; instead we should have used net national product plus the entire existing capital stock. However, then it is not clear that (1) is the correct production function, and empirical implementation demands an accurate K_t series. For these reasons, we did not pursue this route. Instead we took the following alternative. In

place of equation (5) (for which a micro-based justification exists [see Section III]), we posited the ad hoc Solow-type constant-savings rule out of income, along with the conventional assumption that capital depreciates at a constant rate δ . This leads to the following equation for the growth of capital:

$$(5') \quad K_{t+1} = sY_t + (1 - \delta)K_t.$$

Tables 3 and 4 report respectively the unconstrained and constrained (by $\alpha = \frac{1}{3}$) estimates of the parameters when (5') is used in place of (5). Only annual data are used, since we did not have a long enough quarterly time series at the depreciation rates that are commonly used.¹¹ The first set of estimates in Tables 3 and 4 sets δ at 10 percent; the second sets it at 8 percent. The parameter C is the same as before (see footnote 10) with $\gamma = \ln s$. Further detail is in Appendix 3.

All of the estimates imply a significantly *negative* marginal social product of capital, at magnitudes that are simply incredible. Evidently, (5') lends even less support than does (5) to the idea that there are positive externalities to the capital input, or to the notion that in the aggregate production function returns to scale are increasing. The first four tables relied on (5) or (5') to eliminate the capital input from the production function. The next four tables present estimates that use the capital series directly. In Table 5, the low estimate of α is probably due to the high short-run elasticity of output with respect to labor, which is higher than labor's share, $1 - \alpha$. Given that α is set at 0, θ becomes the output elasticity with respect to capital. Thus, Table 5 cannot really be interpreted as supporting the hypothesis that θ is positive. The estimates in Table 6 are more favorable to the hypothesis; α is now constrained to equal $\frac{1}{3}$, yet θ is

¹¹We did experiment with postwar quarterly data, using a 10-year weighted average of past Y_t 's to construct the capital stock. The estimate of θ (with α not constrained) was -1.52 and significantly different from zero. Thus, when (5') is used in place of (5), the annual and quarterly data both yield estimates of θ far below these in Table 1.

TABLE 3—UNCONSTRAINED MAXIMUM-LIKELIHOOD ESTIMATES, YEARLY DATA ONLY, BASED ON AD HOC SAVINGS RULE (5')

δ	Parameter	Estimate	SE	t	P
0.10 ^a	ρ	0.98	0.00	159.52	0.000
	θ	-1.65	0.57	-2.89	0.005
	C	0.29	0.08	3.52	0.000
	λ_1	0.31	0.17	1.83	0.072
	λ_2	0.26	0.16	1.62	0.109
	α	0.02	0.10	0.23	0.818
0.08 ^b	ρ	0.98	0.00	190.49	0.000
	θ	-1.84	0.55	-3.35	0.001
	C	0.33	0.07	4.29	0.000
	λ_1	0.28	0.17	1.59	0.116
	λ_2	0.25	0.16	1.51	0.137
	α	0.02	0.09	0.27	0.784

Note: The estimates are unconstrained in that α need not equal $\frac{1}{3}$.

^aLog likelihood = 101.20 (37 observations, 31 degrees of freedom).

^bLog likelihood = 101.81 (37 observations, 31 degrees of freedom).

TABLE 4—CONSTRAINED MAXIMUM-LIKELIHOOD ESTIMATES, YEARLY DATA ONLY, BASED ON AD HOC SAVINGS RULE (5')

δ	Parameter	Estimate	SE	t	P
0.10 ^a	ρ	0.98	0.00	220.51	0.000
	θ	-2.62	0.57	-4.60	0.000
	C	0.38	0.08	4.62	0.000
	λ_1	0.45	0.19	2.36	0.021
	λ_2	0.40	0.18	2.19	0.032
0.08 ^b	ρ	0.98	0.00	245.34	0.000
	θ	-2.84	0.60	-4.68	0.000
	C	0.41	0.88	4.71	0.000
	λ_1	0.41	0.20	2.07	0.042
	λ_2	0.38	0.18	2.12	0.038

Note: The estimates are constrained in that $\alpha = \frac{1}{3}$.

^aLog likelihood = 96.88 (37 observations, 31 degrees of freedom).

^bLog likelihood = 97.42 (37 observations, 31 degrees of freedom).

TABLE 5—UNCONSTRAINED MAXIMUM-LIKELIHOOD ESTIMATES USING ANNUAL CAPITAL DATA

Parameter	Estimate	SE	t	P
θ	0.31	0.10	2.93	0.006
α	-0.04	0.12	-0.32	0.749
C	-1.16	0.58	-1.97	0.057
λ_1	0.31	0.18	1.70	0.099
λ_2	0.18	0.16	1.09	0.281
ρ	0.83	0.06	12.78	0.000

Notes: The estimates are unconstrained in that α need not equal $\frac{1}{3}$. Log likelihood = 101.68 (37 observations, 31 degrees of freedom).

TABLE 6—CONSTRAINED MAXIMUM-LIKELIHOOD ESTIMATES USING ANNUAL CAPITAL DATA

Parameter	Estimate	SE	<i>t</i>	<i>P</i>
ρ	0.73	0.07	9.66	0.000
θ	0.23	0.05	4.17	0.000
<i>C</i>	-1.44	0.52	-2.75	0.009
λ_1	0.36	0.18	2.03	0.050
λ_2	0.27	0.16	1.68	0.101

Notes: The estimates are constrained in that $\alpha = \frac{1}{3}$. Log likelihood = 97.47 (37 observations, 32 degrees of freedom).

TABLE 7—OLS ESTIMATES USING ANNUAL CAPITAL DATA: LEVELS

Parameter	Coefficient	SE	<i>t</i>	<i>P</i>
<i>C</i>	-0.40	1.42	-0.28	0.777
$1 - \alpha$	-0.13	0.18	-0.70	0.483
$\alpha + \theta$	1.06	0.09	11.67	0.000

Notes: $R^2 = 0.98$, $\bar{R}^2 = 0.98$, residual SS = 0.04, SE of estimate = 0.03, total SS = 4.14, $F_{[3,34]} = 1,420.61$, $P = 0.00$, Durbin-Watson statistic = 0.60 (37 observations, 34 degrees of freedom).

TABLE 8—OLS ESTIMATES USING ANNUAL CAPITAL DATA: GROWTH RATES

Parameter	Coefficient	SE	<i>t</i>	<i>P</i>
<i>C</i>	0.02	0.02	1.26	0.215
$1 - \alpha$	1.01	0.16	6.13	0.000
$\alpha + \theta$	-0.35	0.75	-0.47	0.641

Notes: $R^2 = 0.54$, $\bar{R}^2 = 0.51$, residual SS = 0.01, SE of estimate = 0.01, total SS = 0.02, $F_{[3,33]} = 19.79$, $P = 0.00$, Durbin-Watson statistic = 0.78 (36 observations, 33 degrees of freedom).

still positive and significant. This is the only solid piece of evidence in favor of Romer's hypothesis that we can find in the postwar data. At the same time, the estimate of ρ is surprisingly low. Tables 7 and 8 present ordinary least-squares (OLS) estimates of the aggregate production function. If ρ is close to 1, and if (5) is appropriate for annual (as opposed to quarterly) data, then according to the model in equations (2)–(5) and assumption (6) the OLS estimates of the coefficients in the growth-rates equation (Table 8) are unbiased. On the other hand, the levels equation involves upward bias on the capital coefficient and downward bias on the labor coefficient.

Given the wide diversity of the estimates for θ , α , and ρ reported by the eight tables

in this section, it seems that the assumptions we have added to Romer's model do not substantially improve the ability of his model to rationalize high-frequency data. We therefore agree with Romer's (1987 p. 186) view that data from more countries and longer epochs should provide additional and perhaps better information on the model's parameters, and in particular, on θ . We look at cross-country data next.

III. Cross-Country Evidence on the Univariate Representation for y_t

Under certain assumptions, the Heston-Summers panel data on countries' GNP's will provide additional information on the parameters of the model. This is what we

examine next. We assume that all countries have the same production functions and tastes and that the only differences among them are their initial values for k_t , l_t , and z_t . In particular, z_t obeys the same process in all countries, although its realizations can vary. Because we shall be looking only at the y_t process,¹² and because l_t (in addition to k_t and z_t) will also be treated as unobservable, some further assumptions will now be added. First, we shall assume that $\lambda_1 = \lambda_2 = 0$ in equation (4), so that $\omega_t = \varepsilon_t$. This is done for analytical convenience, and it ought not to make much quantitative difference in this section, where we examine growth rates over a period of 25 years (and not at annual or quarterly growth rates, as was done in the previous section), so that the two-year moving average induced by the λ 's should not matter much, if at all. Second, we shall assume a particular stochastic process for the l_t sequence; in each country l_t is assumed to follow the stochastic process

$$l_t = m + rl_{t-1} + w_t \quad |r| \leq 1$$

where w_t is independently and identically distributed and independent of ε_t . We estimated this equation using U.S. annual data and OLS and obtained

$$l_t = -0.22 + 1.03l_{t-1} \quad R^2 = 0.98 \\ (0.18) \quad (0.02) \quad DW = 1.85.$$

Our maximum-likelihood results together with this suggest that, at least in the U.S., both ρ and r are quite close to unity. We shall then take the bold step of assuming that this is true in all the countries in the Heston-Summers sample.¹³

A. *Gibrat's Law in Growth of GNP*

Although, even under these additional assumptions, a study of the y_t process on its own will not identify θ , it will nevertheless rule out a great many possible values that the pair of crucial parameters (ρ, θ) can assume. One source of information about the behavior of y_t in 115 countries comes from the Heston-Summers sample (which is now updated to 1985). The regression below represents the relationship between the average 1960–1985 rate in GNP growth of a country on the one hand, and its 1960 GNP on the other.¹⁴ That is, the growth of countries is regressed on their initial size. The regression results reveal no significant relation between the two:

$$(8) \quad \Delta y_i = 0.047 - 0.0004 y_i \quad i = 1, \dots, 115 \\ (0.015) \quad (0.001) \\ \text{residual variance} = 0.0004$$

where y_i is the logarithm of 1960 GNP for country i and Δy_i is its growth per year over the 1960–1985 period (standard errors are in parentheses). Thus, the updated sample roughly confirms the nonsignificant relationship between growth and initial size that others have found,¹⁵ a (non-) relation that is in other contexts often referred to as "Gibrat's Law."

To find out what the seeming absence of a relation between size and growth means for our structural parameters, combine equations (2) and (5) to get

$$(9) \quad y_t = (\alpha + \theta)(\gamma + y_{t-1}) + \eta_t$$

¹²The Heston-Summers data set has information on population but not on the labor input. It also has no information on the capital input.

¹³Robert Barro's (1988) cross-country study of the univariate process for log(unemployment) (again, with annual data) revealed some significant cross-country differences in the degree of persistence in that variable. Nevertheless, at least in the postwar samples, the AR(1)-coefficient estimate typically does not differ significantly from unity. There are, however, good reasons to suspect the truth of our assumptions about l_t . First,

human capital should respond positively to ε_t , in much the same way as physical capital. This would tend to induce a positive correlation between l_t and ε_t . On the other hand, fertility responds negatively to income, and this would tend to induce a negative correlation between l_t and longer lags of ε_t .

¹⁴Kuwait was excluded from the regression, as it is an extreme outlier.

¹⁵This finding is for countries as a group, most of which are small and have little R&D investment. For industrialized countries, the result is somewhat different (see William Baumol and Edward Wolff, 1988; Bradford DeLong, 1988).

where $\eta_t \equiv (1 - \alpha)l_t + z_t$. Repeated substitution for lagged y 's leads to the following predicted relation between growth and initial size:

$$(10) \quad y_{t+T} - y_t = [(\alpha + \theta)^T - 1]y_t + \left(\sum_{j=0}^{T-1} (\alpha + \theta)^j \right) \times [(\alpha + \theta)\gamma + \eta_{t+T-j}].$$

The form of equations (9) and (10) depends, of course, on the savings rule in equation (5), and the theoretical justification that we provide for this savings rule rests on the assumption that capital depreciates fully each period. However, this section and the next both look at data at a frequency no greater than once every 25 years, and so this assumed depreciation rate may approximate reality quite well in this context.

Without further work, equation (10) cannot be used to interpret the regression results reported in equation (8), because y_t will be correlated with the disturbance in (10). One can see this by assuming that $r = \rho$, so that $\eta_t = (1 - \alpha)m + \mu + \rho\eta_{t-1} + \varepsilon_t + (1 - \alpha)w_t$, and by recursively substituting for lagged η 's in (7) to obtain

$$(11) \quad \eta_{t+T-j} = \rho^{T-j}\eta_t + (T-j)[\mu + (1 - \alpha)m] + \sum_{s=0}^{T-1-j} \rho^j v_{t+T-j-s}$$

where $v_t \equiv \varepsilon_t + (1 - \alpha)w_t$. As long as $\rho > 0$, innovations in η tend to persist, and η_t and y_t will be positively correlated for each country. Substituting from (11) into (10) then implies that the least-squares estimate of b in the regression $\Delta y_t = a + by_t$ is *identically*

$$(12) \quad \hat{b} = (\alpha + \theta)^T - 1 + \left\{ \left[\text{Cov}_i(\eta_{it}, y_{it}) / \text{Var}_i(y_{it}) \right] \times \sum_{j=0}^{T-1} (\alpha + \theta)^j \rho^{T-j} \right\}.$$

The subscript i on Cov_i and Var_i is there to emphasize that it is i that varies while t is held fixed at $t = 1960$.

To compute the expected value of \hat{b} , we invoke our assumption that the parameters of the y_t process are identical for all countries, in which case the empirical bivariate distribution of (y_{it}, η_{it}) over countries i at t approximates the stationary distribution of (y_t, η_t) for a given country when this distribution exists. When either $\rho \rightarrow 1$, or $(\alpha + \theta) \rightarrow 1$, this distribution blows up,¹⁶ but Appendix 1 shows that the ratio $\text{Cov}(y, \eta) / \text{Var}(y)$ still converges:

$$(13) \quad \lim_{r, \rho \rightarrow 1} [\text{Cov}(\eta_t, y_t) / \text{Var}(y_t)] = 1 - (\alpha + \theta).$$

Substituting from (13) into (12) and noting (once more from Appendix 1) that

$$\lim_{\rho \rightarrow 1} \sum_{j=0}^{T-1} (\alpha + \theta)^j \rho^{T-j} = [1 - (\alpha - \theta)^T] / [1 - (\alpha + \theta)]$$

yields

$$(14) \quad \lim_{\rho \rightarrow 1} E(\hat{b}) = (\alpha + \theta)^T - 1 + \{[1 - (\alpha + \theta)] \times [1 - (\alpha + \theta)^T] / [1 - (\alpha + \theta)]\} = 0.$$

Therefore, if ρ and r are roughly 1, Gibrat's Law will hold *regardless of the value of θ* . This means that the failure of GNP levels to converge does not identify θ .

¹⁶Because y_t acquires a permanent component if $\rho = 1$ or if $\alpha + \theta = 1$. Therefore, the findings of Charles Nelson and Charles Plosser (1982), John Campbell and Gregg Mankiw (1987), and John Cochrane (1988)—that one cannot reject the hypothesis that, in the univariate ARMA representation of GNP, innovations to GNP have a permanent component—do not, by themselves tell us whether $\alpha + \theta = 1$, or whether ρ or r is equal to unity.

As a caveat, we point out that our analysis treats countries as closed economies and looks for scale effects or spillover effects within but not across countries. Yet, geographical borders are in some respects an arbitrary division of geographical space and are therefore "noisy" measures of market areas within which, according to our analysis (and Romer's), these scale or spillover effects are assumed to be confined. Nevertheless, differences in culture and language and the presence of capital controls and other trade barriers do support the use of geographical borders to delineate the extent of the market.¹⁷

B. Growth in the Cross Section and in the Time Series

A second set of questions emerges from our assumption that the bivariate distribution of (y, η) among countries at a point in time is the same as the stationary distribution of (y, η) for a given country over time. The first point to note here is that the truth of this hypothesis is completely independent of the length of the epochs (T_{it}); instead, it has to do with how long the stochastic processes y_{it} have followed the law of motion (9) and with the speed of convergence to the stationary distribution implied by the parameters $\alpha + \theta$, ρ , and r .

The assumption that the cross-section distribution coincides with the stationary distribution can also be tested. Let g_{it} be

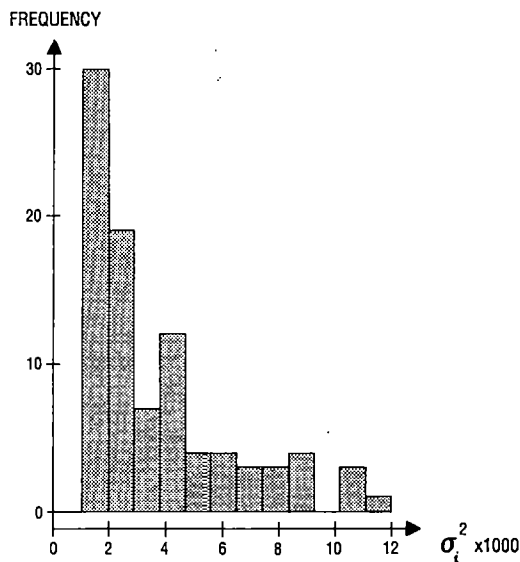


FIGURE 1. THE EMPIRICAL DISTRIBUTION OF COUNTRY-SPECIFIC VARIANCES OF GROWTH

the growth rate of GNP in country i between periods t and $t + 1$. If the hypothesis is true, the distribution of g_{it} ($t = 1960, \dots, 1984$) for each fixed i should be roughly the same as the distribution of g_{it} ($t = 1, \dots, 115$) for each fixed i .¹⁸ In particular, if σ_i^2 and σ_t^2 are the variances of the two respective distributions, we should have $\sigma_i^2 \approx \sigma_t^2$, at least for most i and most t . In fact, σ_i^2 and σ_t^2 both vary considerably as i and t vary, although *on average* they are roughly the same:

$$\left(\frac{1}{115} \right) \sum_{i=1}^{115} \sigma_i^2 = 0.0035$$

$$\left(\frac{1}{25} \right) \sum_{t=1960}^{1984} \sigma_t^2 = 0.0034.$$

The variability of σ_i^2 is documented in Table 9, and the variability of σ_t^2 is described in

¹⁷Of course, if significant cross-country spillovers in knowledge do exist, they surely run mainly from the rich nations toward the poor ones, and if so, they represent a force in support of convergence. In our model, the parameter μ presumably depends inversely on what is known domestically relative to what is known abroad. Two models of learning in situations where different agents know different things are presented in Jovanovic and Rafael Rob (1989) and Jovanovic and Glenn MacDonald (1988). In both of these theoretical models, those who are further behind learn more (through imitation) than those who are closer to the leaders, simply because they have more to learn. This argues for a higher μ for the poorer nations. However, such a perspective ignores the constraints on the capacity of people in a developing country to absorb and apply the technologies that the more advanced countries have already created and put in place (see Raymond Vernon [1989] for a viewpoint that emphasizes these constraints).

¹⁸Actually, this is true only if g_{it} and g_{jt} are independent random variables for $i \neq j$. In fact, Table 9 reveals significant time effects in the world's mean growth rates, especially after the first oil shock. This seems to imply a fair degree of contemporaneous covariation in countries' growth rates from 1974 on. If so, the variance of growth rates over countries at a given date should in fact be less than the variance of growth rates over time in a single country.

TABLE 9—MEAN AND VARIANCE OF THE
CROSS-COUNTRY DISTRIBUTION OF
YEAR-TO-YEAR GROWTH RATES IN GNP

Year	Mean	Variance (σ_t^2)
1961	0.0509	0.0034
1962	0.0528	0.0038
1963	0.0497	0.0023
1964	0.0527	0.0036
1965	0.0498	0.0032
1966	0.0487	0.0034
1967	0.0427	0.0023
1968	0.0536	0.0024
1969	0.0559	0.0030
1970	0.0543	0.0027
1971	0.0493	0.0027
1972	0.0511	0.0029
1973	0.0577	0.0036
1974	0.0429	0.0036
1975	0.0185	0.0059
1976	0.0526	0.0027
1977	0.0398	0.0028
1978	0.0453	0.0030
1979	0.0323	0.0052
1980	0.0327	0.0037
1981	0.0210	0.0061
1982	0.0023	0.0029
1983	0.0133	0.0033
1984	0.0362	0.0046
1985	0.0263	0.0020

Notes: Data are from the Heston-Summers sample, with Kuwait excluded. Growth rates are over the previous year.

the histogram in Figure 1 (in which Iraq, whose $\sigma_t^2 \times 1,000 = 30$, was omitted).¹⁹ A comparison of Table 9 and Figure 1 reveals that the variability of σ_t^2 is greater than that of σ_i^2 , which is what one would expect (if in truth they were equal) on sampling grounds since (a) σ_t^2 averages the variance of g_{it} over 115 countries, while σ_i^2 averages it over 25 years only, and (b) for fixed i , the observations g_{it} are autocorrelated.

C. Medium-Run Differences in Growth Among Countries

While the variability of growth rates among countries does not seem to differ from the variability of growth rates within

¹⁹It turns out that σ_t^2 is significantly negatively correlated with $y_{i,1960}$. That is, initially larger countries have less-variable growth rates. For instance, for the United States, $\sigma_t^2 = 0.0006$.

countries, one might still wonder whether differences in growth rates among countries are too *persistent* to be consistent with our model. The model asserts that, except for initial conditions, the η_t process is the same over countries. One way to pose the question about persistence is to ask about the cross-country variance of the mean growth rates over the 25-year periods. That is, does the model allow for a reasonable chance that some countries will grow much faster than others over a period as long as 25 years, or is this possibility a remote one?

The variance of growth within a country over time ultimately depends on the variance of the process $\eta_t = \varepsilon_t + (1 - \alpha)w_t$. The first step is to see what values of σ_ε^2 and σ_w^2 are compatible with the variance of 25-year annualized growth rates. In Appendix 1, we have calculated the variance of the steady-state distribution of Δy as a function of σ_w^2 and σ_ε^2 . Since $\Delta k_t = \Delta y_{t-1}$, the steady-state variance of Δy coincides with that of Δk , and this expression is provided in Table A1 (Appendix 2), where it is denoted by a_{kk} . If we hypothesize the truth of the Solow neo-classical model and insert $T = 25$, $\theta = 0$, and $\alpha = \frac{1}{3}$ in this expression, it reads $a_{kk} = 24\sigma_w^2 + (48.5)\sigma_\varepsilon^2$. This expression, when divided by 25^2 , should equal the *empirical* value of the cross-country variance of the 25-year averaged growth rates. This value turns out to be equal to slightly less than the variance of the residual in equation (8), namely 0.000355. Thus, setting $a_{kk}/25^2$ equal to this number yields a linear restriction on σ_w^2 and σ_ε^2 , namely

$$(24\sigma_w^2 + 48.5\sigma_\varepsilon^2)/25^2 = 0.000355.$$

This is the restriction on σ_w^2 and σ_ε^2 that will generate the medium-run growth differentials across countries that we observe.

If one were to substitute the U.S. values for σ_w^2 and σ_ε^2 into the above expression, one would obtain an expression for its left-hand side, that is much smaller than the right-hand side. If ρ and r are both unity, one can estimate σ_w^2 and σ_ε^2 by the variance of Δl and Δz , respectively. Doing this for the postwar annual U.S. data yields $\sigma_w^2 =$

0.00044 and $\sigma_e^2 = 0.00037$.²⁰ Substituting these values into the above expression yields just 0.000046, which is too small by a factor of 7.7.

This is not the end of it, however, because of the heteroskedasticity of growth rates over countries. The variance of the U.S. growth rate is 0.0006, whereas the growth-rate variance for the median country is about 0.003, and its mean is 0.0035. Therefore, the variability of growth in the "average" country exceeds that of the United States by a factor of about 5 or 6. Since the U.S. variability underestimates the right-hand side of the above equation by a factor of 7.7, it is likely that, if we had estimates of σ_w^2 and σ_e^2 from the *average* country, we would have explained roughly 65–78 percent of the cross-country variability in growth rates. The discrepancy is therefore far smaller than one would have thought; to account for it, one or more of the parameters that we have assumed to be the same for all countries might have to be made country-specific.²¹

The above arguments suggest that the augmented Solow model with $\theta = 0$ is consistent with the bulk of cross-country growth differentials in the medium run. However, it is clearly not consistent with the tremendous heteroskedasticity in yearly growth rates that Figure 1 highlights, although such

heteroskedasticity could also have been produced by measurement errors with country-specific variances.

While the above discussion leaves some unanswered questions about our model's ability to explain (a) the lack of convergence of GNP *levels* and (b) the existence of persisting differentials in growth *rates*, we should in all fairness point out that an alternative explanation for (a) and (b) simultaneously, is as yet unavailable. For instance, in Romer's (1987) model, with a constant savings propensity tacked on, $\alpha + \theta > 1$ ($\alpha + \theta < 1$) implies that the growth rate will be positively (negatively) correlated with the size of the capital stock, while $\alpha + \theta \approx 1$ implies independence. Under independence, which seems supported by data, differences in growth rates among countries must be due either to differences in technology and savings rates or to shocks. The mere presence of externalities ($\theta > 0$) does not by itself account for differences in growth rates.

IV. The Relation Between Inputs and Output Over Longer Epochs

Consider a regression such as the one that Romer (1987) reports in his equation 18. In country i , over a period length T_{it} , differences in the growth of inputs and outputs are calculated, so that, for instance, $\Delta y_{it} \equiv y_{i,t+T_{it}} - y_{it}$. That is, the regression is

$$(15) \quad T_{it}^{-1} \Delta y_{it} = b + b_k T_{it}^{-1} \Delta k_{it} + b_l T_{it}^{-1} \Delta l_{it} + u_{it}.$$

Romer uses 18 observations that span seven countries (subscript i), and four epochs (subscript t) of at least 30 years in length; the measure of the labor input is hours worked. The least-squares regression results that he reports in his equation 18 are: $\hat{b}_k = 0.87$ with a standard error of 0.08, and $\hat{b}_l = 0.04$ with a standard error of 0.18. Our aim here is to calculate the expectations of \hat{b}_k and \hat{b}_l in light of the added assumptions that we have imposed on the evolution of k , l , and z . The least-squares estimates of the coefficients, denoted by carets, are identi-

²⁰The covariance between Δl and Δz is 0.00012. Ignoring it, as we do here, hurts rather than helps our case. Note also that the estimate of σ_e^2 that we have calculated here is actually consistent with its estimates for the annual data in Tables 1 and 2, because, since we are assuming no moving-average terms in this section, the relevant comparison is with $(1 + \lambda_1^2 + \lambda_2^2)\sigma_e^2$ in those tables.

²¹For instance, the parameter μ might have to be made country-specific. Country-specific fixed effects are, in this context at least, simply a label for one's ignorance, and the calculations about variances reported in the above paragraph are too rough and tentative to convince us that the country-specific fixed effect is needed here. Our z_{it} 's are, we submit, less objectionable, because they at least are stochastically equal among countries, although their particular realizations can vary. Moreover, even if ρ and r are unity, the long-run growth rate of z is just μ for each country, and there can be no long-run differences in growth. We discuss this in greater detail in the next section.

cally equal to²²

$$\begin{bmatrix} \hat{b} \\ \hat{b}_k \\ \hat{b}_l \end{bmatrix} = \begin{bmatrix} 0 \\ \alpha + \theta \\ 1 - \alpha \end{bmatrix} + \begin{bmatrix} n & \bar{k} & \bar{l} \\ \bar{k} & a_{kk} & a_{kl} \\ \bar{l} & a_{kl} & a_{ll} \end{bmatrix}^{-1} \begin{bmatrix} \bar{a} \\ a_{ku} \\ a_{lu} \end{bmatrix}$$

from which one can show that, since $Ea_{lu} = 0$ by equation (6),

$$\begin{aligned} (16) \quad E \begin{bmatrix} \hat{b}_k \\ \hat{b}_l \end{bmatrix} &= \begin{bmatrix} \alpha + \theta \\ 1 - \alpha \end{bmatrix} + \left(E \frac{1}{a_{kk}a_{ll} - a_{kl}^2} \right. \\ &\quad \times \begin{bmatrix} a_{ll} & -a_{kl} \\ -a_{kl} & a_{kk} \end{bmatrix} \begin{bmatrix} a_{ku} \\ 0 \end{bmatrix} \Bigg) \\ &= \begin{bmatrix} \alpha + \theta \\ 1 - \alpha \end{bmatrix} + \left(E \frac{1}{a_{kk}a_{ll} - a_{kl}^2} \right. \\ &\quad \times \begin{bmatrix} a_{ll}a_{ku} \\ -a_{kl}a_{ku} \end{bmatrix} \Bigg). \end{aligned}$$

Since the a_{ij} are all positive, \hat{b}_k will be biased upward while \hat{b}_l will be biased downward, with the bias on \hat{b}_k equalling $-a_{ll}/a_{kl}$ times the bias on \hat{b}_l . Appendix 2 calculates the a_{ij} on the assumption that all countries are subject to the same stochastic process but face different realizations of the ε 's and w 's, as well as different initial condi-

tions. The resulting expressions for the a_{ij} 's are quite messy, but the following limiting results are worth noting.²³

$$(17) \quad \lim_{r, \rho \rightarrow 1} \left\{ \lim_{T \rightarrow \infty} [E(\hat{b}_k)] \right\} = 1$$

$$(18) \quad \lim_{r, \rho \rightarrow 1} \left\{ \lim_{T \rightarrow \infty} [E(\hat{b}_l)] \right\} = 0.$$

These results are of relevance if the epochs (which are of length T) are long and if z and l are roughly random walks, as appears to be the case empirically. Therefore, $E(b_l)$ is zero, regardless of the relative values of σ_ε^2 and σ_w^2 , while $E(\hat{b}_k) = 1$ if $\sigma_w^2 = 0$. These expected values are of course not far from Romer's actual estimates, $\hat{b}_k = 0.87$ and $\hat{b}_l = 0.04$.

Table A1 of Appendix 2 reports the expressions for the a_{ij} that one can use to calculate the bias in the least-squares estimates \hat{b}_k and \hat{b}_l for the cases in which (a) T remains finite but ρ and r tend to unity and (b) ρ and r are less than unity but T goes to infinity. In both cases, the bias remains positive but is difficult to represent analytically in a compact way. The main point is that the limiting values expressed in equations (17) and (18) are good approximations for the values that \hat{b}_k and \hat{b}_l would be expected to take for large T and for r and ρ reasonably close to 1.

Equations (17) and (18) are the same as what Christiano (1987) obtains under a different but related set of assumptions. He allows for country-specific fixed effects μ_i in (3) and m_i in the equation governing the evolution of l , while assuming $\rho = r = 1$ and $\sigma_w^2 = 0$. His theoretical results also assume $\sigma_w^2 = 0$, while his simulations allow for $\sigma_w^2 > 0$; both yield the analogues of (17) and (18), and he argues, as we do, that the results

²²This equation follows directly from the application of the least-squares formula. The number of observations (i.e., the number of country-epoch pairs) is n . The a_{ij} are the raw moments. For instance, $a_{kl} = \sum_{i,t} T_{it}^{-2} \Delta k_{it} \Delta l_{it}$, and so on. The variables with overbars are the mean growth rates over the sample. For instance, $\bar{k} = \sum_{i,t} T_{it}^{-1} \Delta k_{it}$, and so on.

²³We present the results only for this particular limiting case because the general expressions would be very lengthy. Table A1 in Appendix 2 presents results that make it possible to compute $E(\hat{b}_k)$ and $E(\hat{b}_l)$ for finite T or for ρ and r less than unity.

that Romer reports in his equation (18) are consistent with θ being 0.

Several additional insights follow from our analysis, however. In explaining these, it is worthwhile to elaborate on the differences between our model and Christiano's (1987) fixed-effects model. These differences are best explained under the assumption that $\sigma_w^2 = 0$ (i.e., that labor supply is nonrandom). In the fixed-effects model, a country's long-run growth rate is $\mu_i/(1-\alpha)$. In our model, the long-run growth rate for z is

$$\lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T (z_{t+1} - z_t) = \begin{cases} \mu & \text{if } \rho = 1 \\ 0 & \text{if } \rho < 1. \end{cases}$$

Since μ is the same over countries, countries must, in the long run, all grow at the same rate regardless of the value of ρ . Thus, a fixed-effects model delivers a positive variance of long-run growth rates among countries, while ours does not.²⁴ Our view gets further support from recent empirical work by Danny Quah (1990), who argues that, in the Heston-Summers data, income growth rates show convergence to a common value.

In our model, the upward bias on the capital coefficient is positively related to ρ . Continuing with the case in which labor input is nonrandom ($\sigma_w^2 = 0$), we find from the second column of Table A1 that, as T approaches ∞ , the bias on the capital coefficient approaches

$$\frac{\rho[1 - (\alpha + \theta)^2]}{1 + \rho(\alpha + \theta)}.$$

As ρ goes to unity, so that (18) obtains, the bias becomes $1 - (\alpha + \theta)$, while as ρ goes to zero, the bias becomes zero.

The conclusion we draw from this exercise is the same as the one Christiano draws: the regression results that use data on long-run movements of output and both inputs

also provide no support for the hypothesis that θ is significantly positive.

Aside from the hypotheses that Romer, Christiano, and we have advanced, yet another argument can be advanced to rationalize the high estimated coefficient of Δk and the low estimated coefficient of Δl : the growth of the physical capital stock may be strongly positively correlated with growth of labor quality, while growth in hours worked may be negatively related to the growth of labor quality. The latter association is well documented at high frequencies in the micro data, and Romer's (1987) figure 1 provides some indirect evidence for it at lower frequencies for the United States. Since the empirical estimates of equation (15) use hours worked unadjusted for quality, this would tend to cause an upward bias on capital's coefficient and a downward bias on labor's coefficient. This hypothesis, which does not revolve around either external effects or increasing returns, still awaits a careful theoretical treatment, but whatever its eventual fate, it is important to bear in mind that it is independent of the other hypotheses discussed here. Of course, if one were to insist that increases in the capital stock *cause* measured equiproportionate improvements in unmeasured labor quality, then one is back in a framework captured by equation (1), with, say \bar{K}^θ denoting unmeasured labor quality, and this is a framework that the data do not seem to support.

V. Microfoundations of the Model

Our remaining task is to provide a firmer analytic foundation for the equations used in our estimation process. The key element driving the results that stem from the structural equations (2)–(5) and assumption (6) is the dependence of the capital stock, through savings behavior, on the stochastic shock to production in the previous period. If this shock is serially correlated, current output will also depend on the shock in the previous period. Therefore, the correlation of contemporaneous output and capital not only reflects the internal and external impact of contemporaneous capital on output but contains an additional component

²⁴Table A1 contains information about the speed of convergence to zero of variables such as a_{kk} , the variance of the long-run growth rate of capital stock. The table shows that this and other variances and covariances go to zero at the rate T^{-1} when $\rho = 1$, while they go to zero at the rate T^{-2} when $\rho < 1$.

through the joint dependence of output and capital on the previous productivity shock. Ignoring this element results in exaggeration of the importance of capital in production. In this section, we spell out a stochastic overlapping-generations (OLG) model as well as a stochastic Brock-Mirman type of growth model to justify the equations of the previous section, especially equation (5) and assumption (6).

We start with a special OLG model in which the representative agent in generation t faces a wage w_t and a stochastic rate of return on his savings, r_{t+1} . Therefore, his consumption in the second period of his life is $c_{t+1} = (w_t - c_t)r_{t+1}$. We assume that the agent has a logarithmic utility function

$$\beta \ln c_t + (1 - \beta) E \ln (w_t - c_t) r_{t+1}$$

which he maximizes by choosing c_t . The production function is assumed to be of Cobb-Douglas form with a multiplicative productivity shock and is given by (1).

Population growth is stochastic, so that $L_t = L_{t-1}(1 + N_t)$, where N_t is independently and identically distributed with mean zero. The wage rate and the interest rate are equal to the marginal products of labor and capital, respectively. Since the agent bases his saving decision on the marginal product of capital in the next period, he faces a stochastic interest rate (on account both of the stochastic productivity shock and the stochastic growth of labor).

The agent's optimal savings do not depend on the interest rate, so that $s_t = (w_t - c_t) = (1 - \beta)w_t$. Thus, the total savings, which determine next period's capital stock, are

$$\begin{aligned} s_t L_t &= K_{t+1} = (1 - \beta)(1 - \alpha) Z_0 K_t^{\alpha + \theta} L_t^{1 - \alpha} \\ &= (1 - \beta)(1 - \alpha) Y_t \end{aligned}$$

since the share of labor is the fraction $1 - \alpha$ of output. Taking logarithms immediately yields equation (5) and assumption (6) of the previous section.

Before moving on to the infinite-horizon model, we should discuss the role of specific functional forms and assumptions. The log-

arithmic utility function simplifies matters considerably by eliminating the dependence of savings on the next-period rate of return; but its use in this context goes beyond algebraic convenience. Slightly altering the utility function, say to one with a constant relative risk aversion, may yield a savings function that either increases or decreases with the rate of interest, depending on whether the risk-aversion coefficient is less than or greater than unity in absolute value. Since an increase in the productivity shock leads to an expected increase in the shock next period and raises the expected interest rate, productivity shocks may in fact decrease savings and next period's capital stock if the direct wealth effect through wages is dominated, resulting in a *negative* correlation between K_{t+1} and Z_{t+1} and contradicting equation (5) in the previous section. (This issue will also arise in the infinite-horizon model considered below.) In drawing generalizations from the example, therefore, we should keep in mind that we may need a preference specification for which savings are a nondecreasing function of the interest rate.²⁵

We now turn to the specification with an infinitely lived representative agent. Before exploiting specific functional forms, we present a general version to pinpoint again the role of the assumptions embodied in our specific functional forms.

The representative agent has an instantaneous, twice-differentiable utility function $U(c_t, a_t - L_t)$, defined on feasible consumption and leisure sets, where a_t is a stochastic labor endowment and L_t is the labor supply. We may specify a_t as a multiplicative Markov process to reflect population growth, since the actual supply of labor will be endogenously chosen. The twice-differentiable production function is given by

²⁵Another set of problems that plague the OLG model relates to the continuum of equilibria. While our special specification avoids these problems, multiplicities will arise either if outside money is introduced as an additional asset or if the labor-supply decision is endogenized and the logarithmic specification of utility is dropped (for a detailed analysis see Benhabib and Guy Laroque [1988]).

$Y_t = Z_t f(K_t, \bar{K}_t, L_t)$, where Z_t is the stochastic shock to the production function and \bar{K}_t ($\equiv K_t$) enters the production function to reflect an externality. Let δ be the depreciation rate of capital. The agent, facing constraints $K_{t+1} = Z_t f(K_t, \bar{K}_t, L_t) + (1 - \delta)K_t - c_t$ and a given K_0 , maximizes $E \sum_0^\infty \beta^t U(c_t, a_t - L_t)$ by choosing each c_t and L_t after observing Z_t and a_t at every t . In dynamic programming form, the problem can be expressed as

$$V(K_0, Z_0, a_0) = \max_{K_1, L_0} U(Z_0 f(K_0, \bar{K}_0, L_0) + (1 - \delta)K_0 - K_1, a_0 - L_0) + \beta EV(K_1, Z_1, a_1).$$

To simplify matters, we assume that the value function V is twice-differentiable in (K, Z) . (The twice-differentiability of V in certain stochastic cases can be established by the methods of Lawrence Blume and David Easley [1982]). Let the derivatives of $V(K, Z, a)$ with respect to K and Z be denoted by V_K and V_Z and let the second-order derivatives be denoted by V_{KK} , V_{KZ} , and V_{ZZ} . Similarly, let U_c and U_L be the derivatives of the utility function with respect to consumption and leisure, with second derivatives U_{cc} , U_{cL} , and U_{LL} . Again, for simplicity, we will assume that $U_{cL} = 0$. Finally, let the derivatives of the production function be denoted by f_K , $f_{\bar{K}}$, and f_L . Standard methodology establishes the first-order conditions for the representative agent's problem with the usual interpretation:

$$(19) \quad U_c(Z_0 f(K_0, \bar{K}_0, L_0) + (1 - \delta)K_0 - K_1, a_0 - L_0) = \beta V_K(K_1, Z_1, a_1).$$

$$(20) \quad U_c(Z_0 f(K_0, \bar{K}_0, L_0) + (1 - \delta)K_0 - K_1, a_0 - L_0) \times Z_0 f_L(K_0, \bar{K}_0, L_0) = U_L(Z_0 f(K_0, \bar{K}_0, L_0) + (1 - \delta)K_0 - K_1, a_0 - L_0).$$

From equation (20), we can obtain the optimal-labor-supply function as $L_0 = L(K_1, K_0, \bar{K}_0, a_0, Z_0)$. Let L_0^Z and L_0^K indicate the derivatives of L_0 with respect to Z_0 and K_0 .

As discussed earlier, we want to investigate the effect of Z_t on K_{t+1} to establish the nature of the covariance between K_{t+1} and Z_{t+1} . Using (19) and (20),

$$dK_1/dK_0 = [(U_{LL} + U_c Z_0 f_{LL})(U_{cc} Z_0 F_K - U_{cc} Z_0 f_L U_c Z_0 f_{KL})]/D > 0$$

and

$$dK_1/dZ_0 = [(U_{LL} + U_c Z_0 f_{LL}) \times (U_{cc} f - \beta EV_{KZ} dZ_1/dZ_0) - U_{cc} Z_0^2 f_L^2 \beta EV_{KZ} dZ_1/dZ_0]/D$$

where

$$D \equiv (U_{LL} + U_c Z_0 f_{LL})(U_{cc} + \beta EV_{KK}) + \beta EU_{cc} V_{KK} Z_0^2 f_L^2 > 0$$

$$F_K \equiv f_K(K_0, \bar{K}_0, L_0) + f_{\bar{K}}(K_0, \bar{K}_0, L_0) + (1 - \delta)$$

and where V_{KK} and V_{KZ} are evaluated at (K_1, Z_1, a_1) . The policy function $K_1 = h(K_0, Z_0)$ is therefore increasing in K_0 .²⁶ Also, $dK_1/dZ_0 > 0$ if $V_{KZ} > 0$. To evaluate V_{KZ} we first compute

$$V_K(K_0, Z_0, a_0) = U_c Z_0 [f_K + f_{\bar{K}} + (1 - \delta)]$$

²⁶This monotonicity property can be established rigorously without assuming the differentiability of the value function. A proof is in Benhabib and Kazuo Nishimura (1989), in lemma 1 of that of paper's appendix. Although the model there is slightly different, with very minor modifications the proof applies to the present case.

to obtain

$$\begin{aligned} V_{kz}(K_0, Z_0, a_0) \\ = U_c((U_{cc}/U_c)(\partial c_0/\partial Z_0)Z_0 + 1) \\ \times (f_k + f_{\bar{k}} + (1 - \delta)) \\ + U_c Z_0 f_{kL}(dL(K_0, \bar{K}_0, Z_0, a_0)/dZ_0). \end{aligned}$$

The sign of V_{kz} and therefore of dK_1/dZ_0 is ambiguous for the same reasons as in the OLG case. First it depends on the degree of relative risk aversion in the term $(U''/U')(\partial c/\partial Z_0)Z_0 + 1$: if this term is sufficiently negative, $\partial K_1/\partial Z_0$ may become negative. Furthermore, unlike our specification in the OLG model, the labor supply is endogenous. An increase in Z_0 , through its effect on Z_1 , leads to an increase in the expected interest rate and may produce not only lower savings but also a lower labor supply; that is, we may have $dL/dZ < 0$. This also tends to make V_{kz} negative, and if sufficiently strong, may result in $dK_1/dZ_0 < 0$. In the special case of a logarithmic utility function coupled with a Cobb-Douglas production function and full depreciation ($\delta = 1$), V_{kz} is identically zero, as can be easily computed using the solution of this special case, reported below. Therefore, for our purposes, it seems that the main restrictions imposed by a model with logarithmic utility and Cobb-Douglas production with full depreciation are to eliminate the possibility of a saving policy and a labor supply which both decrease in response to increases in the rate of return. To see this, consider the policy function for the general case given by $K_1 = h(K_0, Z_0)$ and assume that $\partial h/\partial Z_0 > 0$. We then have the following lemma.

LEMMA 1: *Let $K_{t+1} = h(K_t, Z_t)$, where h is strictly increasing. If Z_t follows the process described by equations (3) and (4) with $\lambda_i \geq 0$ ($i = 1, 2$), then k_t is stochastically strictly increasing in z_t .*

PROOF:

Recursive substitution for lagged capital shocks in h yields $k_t = \phi(z^t)$, where $z^t \equiv$

$(z_{t-1}, z_{t-2}, \dots)$ and where ϕ is strictly increasing. Applying Bayes' rule along with equation (2) yields that for any vector $\bar{z} \in \mathbb{R}$, $\Pr\{z^t \leq \bar{z} | z_t\}$ is stochastically strictly increasing in z_t , and the claim follows.

A corollary of the lemma is that the steady-state covariance between k_t and z_t is strictly positive, and this is all that is required for an upward bias on the capital coefficient in an ordinary-least-squares context.

The advantage of specifying log utility and Cobb-Douglas production with full depreciation is that we can solve explicitly for the optimal consumption, savings, and labor-supply policies. Using (19) and (20) and adopting the logarithmic instantaneous utility function

$$\lambda \ln c + (1 - \lambda) \ln(a - L)$$

together with the Cobb-Douglas production function $Z_0 K^\alpha \bar{K}^\theta L^{1-\alpha}$, it can easily be verified that savings, or next period's capital stock, will be

$$(21) \quad K_1 = \alpha \beta Z_0 K_0^{\alpha+\theta} L_0^{1-\alpha}$$

and that labor supply is given by

$$(22) \quad L_1 = \frac{\lambda(1-\alpha)a_t}{(1-\lambda)(1-\alpha\beta) + \lambda(1-\alpha)}.$$

If the random endowment follows a multiplicative first-order Markov process, then after taking logarithms, (21) and (22) correspond exactly to equation (5) and assumption (6). Note that, to make labor supply stochastic, we could have made the taste parameter stochastic rather than assume a stochastic endowment. Alternatively, if a , λ , and other relevant parameters in (22) were constant, labor supply would be as well, and we would run into identification problems in the previous section. (Note that, in the general specification of the model, labor would be stochastic even if a and λ were fixed.)

We conclude, therefore, that the specifications represented by equation (5) and assumption (6), which drive our results in the

previous section and which underlie our empirical conclusions, can be obtained under reasonable assumptions in either the OLG or the infinitely-lived-agent models of stochastic growth.

VI. Conclusions

Given our assumption that knowledge causes capital but not the other way around, our failure to find a positive θ implies nothing whatsoever about externalities in the generation of knowledge. The Solow model with no externalities to either labor or capital but with stochastic shocks to knowledge does not appear to be contradicted by long-run data on output and the two inputs; furthermore, it is also consistent with micro evidence on knowledge spillovers. The apparent validity of Gibrat's law in countries' GNP series does not contradict it, nor do the seemingly sizable medium-run differentials in growth rates over countries. Moreover, the model fits in with the recent business-cycle literature that explains properties of cycles with productivity shocks.

The realizations of our technology shocks, the z 's, are allowed to differ over countries, but the stochastic process forming them is assumed to be the same, as indeed are all the parameters of our model. That technology shocks can assume different values over countries seems reasonable if one interprets these shocks broadly to include shifts in institutional and organizational structures, such as shifts in the corporate, legal, or bureaucratic structures, or even in attitudes toward work. These elements can greatly enhance or retard the effective use and operation of factors of production. While such changes in institutional or organizational structures may not be permanent, they tend to be quite persistent, so that productivity in different economies can diverge for extended periods of time.

No doubt, a quantum leap in our understanding of growth will occur only when the engine of growth, namely the z_t process, is successfully endogenized. What we think we have shown here, however, is that this engine is fueled primarily by something other than physical capital.

APPENDIX

Appendix 1: The Derivation of Equation (13)

Note first that

$$\sum_{j=0}^{T-1} (\alpha + \theta)^j \rho^{T-j} = \rho^T \frac{1 - [\alpha + \theta/\rho]^T}{1 - (\alpha + \theta)/\rho}.$$

If $(\alpha + \theta) = \rho$, this expression is equal to $T\rho^T$. Next, note from equations (2) and (5) that $k_{t+1} = \gamma + (\alpha + \theta)k_t + \eta_t$, where $\eta_t = (1 - \alpha)l_t + z_t$. Then,

$$\begin{aligned} \text{Cov}(\eta_t, k_t) &= \text{Cov}(\eta_t, (\alpha + \theta)k_{t-1} + \eta_{t-1}) \\ &= (\alpha + \theta)\text{Cov}(\eta_t, k_{t-1}) \\ &\quad + \text{Cov}(\eta_t, \eta_{t-1}). \end{aligned}$$

Expanding further, we obtain

$$\begin{aligned} \text{(A1)} \quad \text{Cov}(\eta_t, k_t) &= \sum_{j=1}^{\infty} (\alpha + \theta)^{j-1} \text{Cov}(\eta_t, \eta_{t-j}). \end{aligned}$$

Since $(1 - \alpha)l_t = (1 - \alpha)m + (1 - \alpha)l_{t-1} + (1 - \alpha)w_t$, then if $\rho = r$,

$$\begin{aligned} \eta_t &= (1 - \alpha)m + \mu + \rho\eta_{t-1} \\ &\quad + (\varepsilon_t + (1 - \alpha)w_t) \end{aligned}$$

so that $\text{Cov}(\eta_t, \eta_{t-j}) = \rho^j \sigma_{\eta}^2$, where $\sigma_{\eta}^2 = [\sigma_{\varepsilon}^2 + (1 - \alpha)^2 \sigma_w^2] / (1 - \rho^2)$. Then, using (A1),

$$\begin{aligned} \text{(A2)} \quad \text{Cov}(\eta_t, k_t) &= \sigma_{\eta}^2 \sum_{j=1}^{\infty} \rho^j (\alpha + \theta)^{j-1} \\ &= \rho \sigma_{\eta}^2 / [1 - \rho(\alpha + \theta)]. \end{aligned}$$

From equation (2),

$$\text{Cov}(\eta_t, y_t) = (\alpha + \theta)\text{Cov}(\eta_t, k_t) + \sigma_{\eta}^2.$$

Substituting into this expression from (A2) or yields

$$\begin{aligned} (A3) \quad \text{Cov}(\eta_t, y_t) &= \sigma_\eta^2 \{1 + \rho(\alpha + \theta) / [1 - \rho(\alpha + \theta)]\} \\ &= \sigma_\eta^2 / [1 - \rho(\alpha + \theta)]. \end{aligned}$$

Next, we need to compute $\text{Var}(y_t)$. Since $y_t = (\alpha + \theta)k_t + \eta_t$,

$$\begin{aligned} (A4) \quad \text{Var}(y_t) &= (\alpha + \theta)^2 \text{Var}(k_t) + \sigma_\eta^2 \\ &\quad + 2(\alpha + \theta) \text{Cov}(\eta_t, k_t). \end{aligned}$$

Now, since $k_{t+1} = \gamma + y_t$, $\text{Var}(y_t) = \text{Var}(k_t)$. Using this in (A4) and substituting from (A3) into (A4) for $\text{Cov}(\eta_t, k_t)$ yields

$$\begin{aligned} (A5) \quad \text{Var}(y_t) &= \left(\frac{\sigma_\eta^2}{1 - (\alpha + \theta)^2} \right) \left(1 + \frac{2(\alpha + \theta)\rho}{1 - \rho(\alpha + \theta)} \right). \end{aligned}$$

The expressions in (A4) and (A5) both explode when ρ approaches 1, because σ_η^2 goes to infinity, but their ratio does not:

$$\lim_{\rho \rightarrow 1} [\text{Cov}(\eta, y) / \text{Var}(y)] = 1 - \alpha - \theta.$$

This is equation (13) of the text, since, by assumption, $\rho = r$.

Appendix 2

Here, we derive expressions for the a_{ij} in equation (16) under various assumptions. Deterministic components of z and l are ignored. We assume in equation (4) that $\lambda_1 = \lambda_2 = 0$, so that $\varepsilon_t = w_t$ and so that the w_t are also independently and identically distributed. Repeated substitution in (3) leads to

$$\begin{aligned} z_{t+T} &= \rho^T z_t + \rho^{T-1} \varepsilon_t \\ &\quad + \rho^{T-2} \varepsilon_{t+1} + \dots + \varepsilon_{t+T-1} \end{aligned}$$

$$\begin{aligned} \Delta^T z_t &\equiv z_{t+T} - z_t \\ &= (\rho^T - 1) z_t + \rho^{T-1} \varepsilon_t \\ &\quad + \rho^{T-2} \varepsilon_{t+1} + \dots + \varepsilon_{t+T-1}. \end{aligned}$$

However,

$$\begin{aligned} z_t &= \rho^{j+1} z_{t-j-1} + \rho^j \varepsilon_{t-j-1} + \rho^{j-1} \varepsilon_{t-j} \\ &\quad + \rho^{j-2} \varepsilon_{t-j+1} + \dots + \varepsilon_{t-1} \end{aligned}$$

so that

$$\begin{aligned} \Delta^T z_t &= (\rho^T - 1) \\ &\quad \times (\rho^{j+1} z_{t-j-1} + \rho^j \varepsilon_{t-j-1} + \dots + \varepsilon_{t-1}) \\ &\quad + \rho^{T-1} \varepsilon_t + \rho^{T-2} \varepsilon_{t+1} + \dots + \varepsilon_{t+T-1} \end{aligned}$$

and also

$$\begin{aligned} \Delta^T z_{t-j-1} &= (\rho^T - 1) z_{t-j-1} \\ &\quad + \rho^{T-1} \varepsilon_{t-j-1} + \rho^{T-2} \varepsilon_{t-j} \\ &\quad + \rho^{T-3} \varepsilon_{t-j+1} + \dots + \varepsilon_{t-j+T-2}. \end{aligned}$$

Note that subscripts on $\Delta^T z_t$ for the ε 's run from $t-j-1$ to $t-1+T$ and that subscripts on $\Delta^T z_{t-j-1}$ for the ε 's run from $t-j-1$ to $t-1+T-j-1$. We shall consider two separate cases: (i) $T-j-1 > 0$ and (ii) $T-j-1 \leq 0$, both for $j \geq 0$.

Case (i). For this case,

$$\begin{aligned} \text{Cov}(\Delta^T z_t, \Delta^T z_{t-j-1}) &= [\sigma_\varepsilon^2 / (1 - \rho^2)] [(\rho^T - 1)^2 \rho^{j+1}] \\ &\quad + \sigma_\varepsilon^2 (\rho^T - 1) \sum_{i=1}^{j+1} \rho^{j+1-i} \rho^{T-i} \\ &\quad + \sigma_\varepsilon^2 \sum_{i=1}^{T-j-1} \rho^{T-i} \rho^{t-j-1-i} \end{aligned}$$

$$\begin{aligned}
&= [\sigma_\varepsilon^2 / (1 - \rho^2)] (\rho^T - 1)^2 \rho^{j+1} \\
&\quad + \sigma_\varepsilon^2 (\rho^T - 1) \rho^{T+j-1} \\
&\quad \times [(1 - \rho^{-2(j+1)}) / (1 - \rho^{-2})] \\
&\quad + \sigma_\varepsilon^2 \rho^{2T-j-3} \\
&\quad \times [(1 - \rho^{-2(T-j-1)}) / (1 - \rho^{-2})].
\end{aligned}$$

As a check on the algebra, note that $\lim_{\rho \rightarrow 1} [\text{Cov}(\Delta^T z_t, \Delta^T z_{t-j-1})] = \sigma_\varepsilon^2 (T - j - 1)$, because the first term goes to zero by L'Hôpital's rule and the second term is zero. This result is exactly as expected.

Case (ii). For $(T - j - 1) \leq 0$,

$$\begin{aligned}
&\text{Cov}(\Delta^T z_t, \Delta^T z_{t-j-1}) \\
&= [\sigma_\varepsilon^2 / (1 - \rho^2)] (\rho^T - 1)^2 \rho^{j+1} \\
&\quad + (\rho^T - 1) \left(\sum_{i=1}^T \rho^{j+1-i} \rho^{T-i} \right) \sigma_\varepsilon^2.
\end{aligned}$$

Again, as a check on the algebra, note that $\lim_{\rho \rightarrow 1} [\text{Cov}(\Delta^T z_t, \Delta^T z_{t-j-1})] = 0$, as it should. As a further check, note that when $j = 0$, we have

$$\begin{aligned}
\text{Var}(\Delta^T z_t) &= [\sigma_\varepsilon^2 / (1 - \rho^2)] (\rho^T - 1)^2 \\
&\quad + [(\rho^{T-2} - \rho^{-2}) / (1 - \rho^{-2})] \sigma_\varepsilon^2
\end{aligned}$$

so that $\lim_{\rho \rightarrow 1} [\text{Var}(\Delta^T z_t)] = \sigma_\varepsilon^2 T$, as it should. Moreover, for $\rho < 1$,

$$\begin{aligned}
&\lim_{T \rightarrow \infty} [\text{Var}(\Delta^T z_t)] \\
&= \sigma_\varepsilon^2 / (1 - \rho^2) \\
&\quad + [-\rho^{-2} / (1 - \rho^{-2})] \sigma_\varepsilon^2 \\
&= 2\sigma_\varepsilon^2 / (1 - \rho^2).
\end{aligned}$$

Now we shall compute $\text{Cov}(\Delta^T k_t, \Delta^T z_t)$, first for arbitrary ρ and T , and then we shall

take limits. Combining cases (i) and (ii),

$$\begin{aligned}
&\text{Cov}(\Delta^T z_t, \Delta^T k_t) \\
&= \sum_{j=0}^{\infty} (\alpha + \theta)^j \text{Cov}(\Delta^T z_t, \Delta^T z_{t-j-1}) \\
&= \sum_{j=0}^{T-2} (\alpha + \theta)^j \text{Cov}(\Delta^T z_t, \Delta^T z_{t-j-1}) \\
&\quad + \sum_{j=T-1}^{\infty} (\alpha + \theta)^j \\
&\quad \times \text{Cov}(\Delta^T z_t, \Delta^T z_{t-j-1}) \\
&= \sum_{j=0}^{T-2} (\alpha + \theta)^j \\
&\quad \times \left\{ [\sigma_\varepsilon^2 / (1 - \rho^2)] (\rho^T - 1)^2 \rho^{j+1} \right. \\
&\quad \quad + \sigma_\varepsilon^2 (\rho^T - 1) \sum_{i=1}^{j+1} \rho^{j+1-i} \rho^{T-i} \\
&\quad \quad \left. + \sigma_\varepsilon^2 \sum_{i=1}^{T-j-1} \rho^{T-i} \rho^{T-j-1-i} \right\} \\
&\quad + \sum_{j=T-1}^{\infty} (\alpha + \theta)^j \\
&\quad \times \left\{ [\sigma_\varepsilon^2 / (1 - \rho^2)] (\rho^T - 1)^2 \rho^{j+1} \right. \\
&\quad \quad \left. + \sigma_\varepsilon^2 (\rho^T - 1) \sum_{i=1}^T \rho^{j+1-i} \rho^{T-i} \right\}.
\end{aligned}$$

If we now let T approach ∞ , so that the second summation on the right goes to zero,

we obtain

$$\begin{aligned}
 & \sum_{j=0}^{\infty} (\alpha + \theta)^j \left\{ \left[\sigma_{\varepsilon}^2 / (1 - \rho^2) \right] \left[(\rho^T - 1)^2 \rho^{j+1} \right] \right. \\
 & \quad + (\rho^T - 1) \rho^{T-1} \rho^j \\
 & \quad \times \left[(1 - \rho^{-2(j+1)}) / (1 - \rho^{-2}) \right] \sigma_{\varepsilon}^2 \\
 & \quad + \sigma_{\varepsilon}^2 \rho^{2T-3} \rho^{-j} \\
 & \quad \times \left. \left[(1 - \rho^{-2(T-j-1)}) / (1 - \rho^{-2}) \right] \right\} \\
 & = (\rho^T - 1)^2 \left[\sigma_{\varepsilon}^2 / (1 - \rho^2) \right] \\
 & \quad \times \rho [1 - (\alpha + \theta) \rho]^{-1} \\
 & \quad + \left[\sigma_{\varepsilon}^2 / (1 - \rho^2) \right] (\rho^T - 1) \rho^{T-1} \\
 & \quad \times \left\{ [1 - (\alpha + \theta) \rho]^{-1} \right. \\
 & \quad \quad \left. - \rho^{-2} [1 - (\alpha + \theta) \rho^{-1}]^{-1} \right\} \\
 & \quad + \rho^{2T-3} \left\{ [1 - (\alpha + \theta) \rho - 1]^{-1} \right. \\
 & \quad \quad \left. - \rho^{-1} / [1 - (\alpha + \theta) \rho] \right\} \sigma_{\varepsilon}^2 \\
 & \quad \times (1 - \rho^{-2})^{-1}.
 \end{aligned}$$

Now we note that as T approaches ∞ , the second term above also goes to zero. The first term goes to $\sigma_{\varepsilon}^2 \rho / (1 - \rho^2) [1 - \rho(\alpha + \theta)]$, while the third term goes to $\rho^{-1} \sigma_{\varepsilon}^2 / (1 - \rho^{-2}) [1 - \rho(\alpha + \theta)]$. Therefore,

$$\begin{aligned}
 \text{(A5a)} \quad & \lim_{T \rightarrow \infty} [\text{Cov}(\Delta^T k_t, \Delta^T z_t)] \\
 & = \{ \sigma_{\varepsilon}^2 / [1 - \rho(\alpha + \theta)] \} \\
 & \quad \times [\rho / (1 - \rho^2) \\
 & \quad \quad - \rho^{-1} / (1 - \rho^{-2})] \\
 & = 2\sigma_{\varepsilon}^2 \rho / (1 - \rho^2) [1 - \rho(\alpha + \theta)].
 \end{aligned}$$

Next we calculate the limit as ρ approaches 1, for fixed T .

$$\begin{aligned}
 \text{(A5b)} \quad & \lim_{\rho \rightarrow 1} [\text{Cov}(\Delta^T z_t, \Delta^T k_t)] \\
 & = \sigma_{\varepsilon}^2 \sum_{j=0}^{T-2} (\alpha + \theta)^j (T - j - 1) \\
 & = \sigma_{\varepsilon}^2 [T - 1 + (\alpha + \theta) \\
 & \quad \times (T + 2) \dots (\alpha + \theta)^{T-2}] \\
 & = \sigma_{\varepsilon}^2 \{ T [1 - (\alpha + \theta)]^{-1} \\
 & \quad - [1 - (\alpha + \theta)^T] \\
 & \quad \times [1 - (\alpha + \theta)]^{-2} \} \\
 & = \sigma_{\varepsilon}^2 \{ T [1 - (\alpha + \theta)] \\
 & \quad - [1 - (\alpha + \theta)^T] \} \\
 & \quad \times [1 - (\alpha + \theta)]^{-2}.
 \end{aligned}$$

Next we turn to the computation of $\text{Var}(\Delta^T k_t)$. We have

$$\begin{aligned}
 \text{Var}(\Delta^T k_t) & = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} (\alpha + \theta)^{i+j} \\
 & \quad \times \left[(1 - \alpha)^2 \text{Cov}(\Delta^T l_{t-j}, \Delta^T l_{t-i}) \right. \\
 & \quad \left. + \text{Cov}(\Delta^T z_{t-j}, \Delta^T z_{t-i}) \right].
 \end{aligned}$$

Let

$$\begin{aligned}
 \hat{A} & = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} (\alpha + \theta)^{i+j} (1 - \alpha)^2 \\
 & \quad \times \text{Cov}(\Delta^T l_{t-j}, \Delta^T l_{t-i}).
 \end{aligned}$$

We will compute \hat{A} later. If we let ρ approach 1 for a given T , remembering that

$$\lim_{\rho \rightarrow 1} [\text{Cov}(\Delta^T z_t, \Delta^T z_{t-j-1})] = 0$$

for $t - j - 1 \leq 0$, we obtain

$$\begin{aligned} \text{Var}(\Delta^T k_t) &= \sum_{j=0}^{\infty} \sum_{i=j+1}^{T+j} (\alpha + \theta)^{i+j} \\ &\quad \times [T - (i - j)] \sigma_\varepsilon^2 \\ &\quad + \sum_{i=0}^{\infty} \sum_{j=i+1}^{T+i} (\alpha + \theta)^{i+j} \\ &\quad \times [T - (j - i)] \sigma_\varepsilon^2 \\ &\quad + \sum_{k=0}^{\infty} (\alpha + \theta)^{2k} T \sigma_\varepsilon^2 + \hat{A} \\ &= \left(2\sigma_\varepsilon^2 (\alpha + \theta) \left\{ T / [1 - (\alpha + \theta)] \right. \right. \\ &\quad \left. \left. - [1 - (\alpha + \theta)^T] / [1 - (\alpha + \theta)]^2 \right\} \right. \\ &\quad \left. + T \sigma_\varepsilon^2 [1 - (\alpha + \theta)^2]^{-1} \right) \\ &\quad \times [1 - (\alpha + \theta)^2]^{-1} + \hat{A}. \end{aligned}$$

We also compute $\text{Var}(\Delta^T k_t)$ for $\rho < 1$ as T approaches ∞ ; we have

$$\begin{aligned} \text{Var}(\Delta^T k_t) &= (1 - \alpha)^2 \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \text{Cov}(\Delta^T l_{t-j}, \Delta^T l_{t-i}) \\ &\quad \times (\alpha + \theta)^{i+j} \\ &\quad + \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \text{Cov}(\Delta^T z_{t-j}, \Delta^T z_{t-i}) \\ &\quad \times (\alpha + \theta)^{i+j}. \end{aligned}$$

Let \hat{A} again denote the first of these expressions. We shall compute it later. Next,

observe that, for $i > j$ and $T - (i - j) > 0$,

$$\begin{aligned} \text{Cov}(\Delta^T z_{t-i}, \Delta^T z_{t-j}) &= [\sigma_\varepsilon^2 / (1 - \rho^2)] \\ &\quad \times [(\rho^j - 1)^2 \rho^{i-j} \\ &\quad + \sigma_\varepsilon^2 (\rho^T - 1) \rho^{T+(i-j)-2} \\ &\quad \times (1 - \rho^{-2(i-j)}) / (1 - \rho^{-2})] \\ &\quad + \sigma_\varepsilon^2 \rho^{2T-(i-j)-2} \\ &\quad \times (1 - \rho^{-2T+2(i-j)}) / (1 - \rho^{-2}). \end{aligned}$$

If we let T approach ∞ , note that $T - (i - j) > 0$ and $T - (j - i) > 0$ for all fixed i, j . Now, break the summation for $\text{Var}(\Delta^T k_t)$ into three parts: $i > j$, $i < j$, and $i = j$. The expressions for $i > j$ and $i < j$ are symmetric, so compute twice the value for $i > j$:

$$\begin{aligned} \lim_{T \rightarrow \infty} [\text{Var}(\Delta^T k_t)] &= \lim_{T \rightarrow \infty} \left(2 \sum_{j=0}^{\infty} \sum_{i=j+1}^{\infty} (\alpha + \theta)^{i+j} \right. \\ &\quad \times [\sigma_\varepsilon^2 / (1 - \rho^2)] \\ &\quad \times \{ (\rho^T - 1)^2 \rho^{i-j} \\ &\quad + (\rho^T - 1) [\sigma_\varepsilon^2 / (1 - \rho^{-2})] \\ &\quad \times \rho^{T-2} \rho^{i-j} (1 - \rho^{-2i} \rho^{2j}) \\ &\quad + [\sigma_\varepsilon^2 / (1 - \rho^{-2})] \\ &\quad \times [\rho^{2T-2} \rho^{-i} \rho^j (1 - \rho^{-2T} \rho^{2i} \rho^{-2j})] \} \\ &\quad \left. + \sum_{j=0}^{\infty} (\alpha + \theta)^{2j} \right. \\ &\quad \times \{ [\sigma_\varepsilon^2 / (1 - \rho^2)] (\rho^T - 1)^2 \\ &\quad + [\sigma_\varepsilon^2 / (1 - \rho^{-2})] \\ &\quad \times (\rho^T - 1) \rho^{T-2} (0) \\ &\quad + [\sigma_\varepsilon^2 / (1 - \rho^{-2})] \\ &\quad \times [\rho^{2T-2} (1 - \rho^{-2T})] \} + \hat{A} \Big) \end{aligned}$$

$$\begin{aligned}
&= \lim_{T \rightarrow \infty} \left(2\sigma_\varepsilon^2 \left[\sum_{j=0}^{\infty} (\rho^T - 1)^2 (\alpha + \theta)^j \right. \right. \\
&\quad \times \rho^{-j} (\alpha + \theta)^{j+1} \rho^{j+1} \\
&\quad \times [1 - (\alpha + \theta)\rho]^{-1} (1 - \rho) \\
&\quad + \rho^{T-1} (1 - \rho^{-2})^{-1} \\
&\quad \times (\alpha + \theta)^j \rho^{T-2} \rho^{-j} \\
&\quad \times \{\rho^{j+1} (\alpha + \theta)^{j+1} \\
&\quad \times [1 - (\alpha + \theta)\rho]^{-1} \\
&\quad - \rho^{2j} (\alpha + \theta)^{j+1} \rho^{-(j+1)} \\
&\quad \times [1 - (\alpha + \theta)\rho^{-1}]^{-1}\} \\
&\quad + (1 - \rho^{-2}) \rho^{2T-2} (\alpha + \theta)^j \\
&\quad \times \rho^j \{ (\alpha + \theta)^{j+1} \rho^{-(j+1)} \\
&\quad \times [1 - (\alpha + \theta)\rho^{-1}]^{-1} \\
&\quad - \rho^{-2T} \rho^{-2j} (\alpha + \theta)^{j+1} \\
&\quad \times \rho^{j+1} [1 - (\alpha + \theta)\rho]^{-1} \} \\
&\quad + [\sigma_\varepsilon^2 (1 - \rho^{-2})^{-1} (\rho^T - 1)^2 \\
&\quad + \sigma_\varepsilon^2 (1 - \rho^{-1})^{-1} \rho^{2T-2} \\
&\quad - \sigma_\varepsilon^2 (1 - \rho^{-2})^{-1} \rho^{-2}] \\
&\quad \times [1 - (\alpha + \theta)^2]^{-1} + \hat{A} \Big) \\
&= \lim_{T \rightarrow \infty} \left(2 \{ [\sigma_\varepsilon^2 / (1 - \rho^2)] (\rho^T - 1)^2 \right. \\
&\quad \times \rho (\alpha + \theta) [1 - (\alpha + \theta)^2] \\
&\quad \times [1 - (\alpha + \theta)\rho]^{-1} \\
&\quad + (\rho^T - 1) \rho^{T-2} [\sigma_\varepsilon^2 / (1 - \rho^{-2})] \\
&\quad \times \rho (\alpha + \theta) \\
&\quad \times [1 - (\alpha + \theta)^2]^{-1} [1 - (\alpha + \theta)\rho]^{-1} \\
&\quad - [\sigma_\varepsilon^2 / (1 - \rho^{-2})] (\rho^T - 1) \\
&\quad \times \rho^{T-2} (\alpha + \theta) \rho^{-1} [1 - (\alpha + \theta)^2]^{-1} \\
&\quad \times [1 - (\alpha + \theta)\rho^{-1}]^{-1} \\
&\quad + [\sigma_\varepsilon^2 / (1 - \rho^{-2})] \rho^{2T-2} \\
&\quad \times (\alpha + \theta) \rho^{-1} [1 - (\alpha + \theta)^2]^{-1} \\
&\quad \times [1 - (\alpha + \theta)\rho^{-1}]^{-1} \\
&\quad - [\sigma_\varepsilon^2 / (1 - \rho^{-2})] (\alpha + \theta) \\
&\quad \times \rho^{-1} [1 - (\alpha + \theta)^2]^{-1} \\
&\quad \times [1 - (\alpha + \theta)\rho^{-1}]^{-1} \} \\
&\quad + [1 - (\alpha + \theta)^2]^{-1} \\
&\quad \times [\sigma_\varepsilon^2 (1 - \rho^2)^{-1} (\rho^T - 1)^2 \\
&\quad + \sigma_\varepsilon^2 (1 - \rho^{-2})^{-1} \rho^{2T-2} \\
&\quad - \sigma_\varepsilon^2 (1 - \rho^{-2})^{-1} \rho^{-2}] + \hat{A} \Big) \\
&= \lim_{T \rightarrow \infty} \left([\sigma_\varepsilon^2 / (1 - \rho^2)] [1 - (\alpha + \theta)^2]^{-1} \right. \\
&\quad \times \{ 2(\rho^T - 1)^2 [1 - \rho(\alpha + \theta)]^{-1} \\
&\quad \times \rho (\alpha + \theta) \\
&\quad - 2(\rho^T - 1) \rho^T \rho \\
&\quad \times (\alpha + \theta) [1 - (\alpha + \theta)\rho]^{-1} \\
&\quad + 2(\rho^T - 1) \rho^T (\alpha + \theta) \\
&\quad \times \rho^{-1} [1 - (\alpha + \theta)\rho^{-1}] \\
&\quad - 2\rho^{2T} (\alpha + \theta) \rho^{-1} \\
&\quad \times [1 - (\alpha + \theta)\rho^{-1}]^{-1} \\
&\quad + 2(\alpha + \theta)\rho \\
&\quad \times [1 - (\alpha + \theta)\rho]^{-1} \\
&\quad + [(\rho^T - 1)^2 - \rho^{2T} + 1] \\
&\quad + 2(1 - \rho^T) \} + \hat{A} \Big)
\end{aligned}$$

$$\begin{aligned}
&= \lim_{T \rightarrow \infty} \left(2[\sigma_\varepsilon^2 / (1 - \rho^2)] \right. \\
&\quad \times [1 - (\alpha + \theta)^2] \\
&\quad \times \left[(\rho^T - 1)^2 \rho(\alpha + \theta) \right. \\
&\quad \times [1 - (\alpha + \theta)\rho]^{-1} + (\rho^T - 1)\rho^T \\
&\quad \times \left\{ \rho^{-1} [1 - (\alpha + \theta)\rho^{-1}]^{-1} \right. \\
&\quad \quad \left. - \rho [1 - (\alpha + \theta)\rho]^{-1} \right\} (\alpha + \theta) \\
&\quad \left. + (\alpha + \theta) \left\{ \rho [1 - (\alpha + \theta)\rho]^{-1} \right. \right. \\
&\quad \quad \left. \left. - \rho^{-1} [1 - (\alpha + \theta)\rho^{-1}] \rho^{2T} \right\} \right. \\
&\quad \left. + 1 - \rho^T \right] + \hat{A} \Big).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{(A6)} \quad &\lim_{T \rightarrow \infty} [\text{Var}(\Delta^T k_t)] \\
&= 2[\sigma_\varepsilon^2 / (1 - \rho^2)] \\
&\quad \times [1 - (\alpha + \theta)^2]^{-1} \\
&\quad \times \{ 2(\alpha + \theta)\rho \\
&\quad \quad \times [1 - (\alpha + \theta)\rho]^{-1} + 1 \} \\
&\quad + \lim_{T \rightarrow \infty} \hat{A}.
\end{aligned}$$

Finally, we need to compute \hat{A} . Since $l_t = rl_{t-1} + w_t$, the process l_t behaves like the z_t process, with r replacing ρ and with w replacing ε . Therefore, using earlier formulas for z ,

$$\begin{aligned}
\text{(A7)} \quad &\text{Cov}(\Delta^T l_{t-i}, \Delta^T l_{t-j}) \\
&= [T - (i - j)] \sigma_w^2 \quad \text{for } r \rightarrow 1
\end{aligned}$$

so that $\text{Var}(\Delta^T l_t) \rightarrow T\sigma_w^2$ and

$$\begin{aligned}
&\text{Cov}(\Delta^T l_{t-i}, \Delta^T l_{t-j}) \\
&= 2[\sigma_w^2 / (1 - r^2)] r^{i-j} \quad \text{for } T \rightarrow \infty, r < 1
\end{aligned}$$

so that $\text{Var}(\Delta^T l_t) \rightarrow 2\sigma_w^2 / (1 - r^2)$.

Now, for $r = 1$,

$$\begin{aligned}
\hat{A} &= (1 - \alpha)^2 \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \text{Cov}(\Delta^T l_{t-i}, \Delta^T l_{t-j}) \\
&\quad \times (\alpha + \theta)^{i+j} \\
&= \{ (1 - \alpha)^2 / [1 - (\alpha + \theta)^2] \} \\
&\quad \times \{ 2\sigma_w^2(\alpha + \theta) \{ T / [1 - (\alpha + \theta)] \\
&\quad \quad - [1 - (\alpha + \theta)^T] \\
&\quad \quad \times [1 - (\alpha + \theta)]^{-2} \} \\
&\quad + T\sigma_w^2 \}.
\end{aligned}$$

On the other hand, for $r < 1$,

$$\begin{aligned}
&\lim_{T \rightarrow \infty} \hat{A} \\
&= \lim_{T \rightarrow \infty} \left((1 - \alpha)^2 \right. \\
&\quad \times \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \text{Cov}(\Delta^T l_{t-i}, \Delta^T l_{t-j}) \\
&\quad \times (\alpha + \theta)^{i+j} \Big) \\
&= \lim_{T \rightarrow \infty} \left((1 - \alpha)^2 [2\sigma_w^2 / (1 - r^2)] \right. \\
&\quad \times [1 - (\alpha + \theta)^2] \\
&\quad \times \left\{ (r^T - 1)^2 r(\alpha + \theta) \right. \\
&\quad \quad \times [1 - (\alpha + \theta)r]^{-1} \\
&\quad \quad + (r^T - 1)r^T \\
&\quad \quad \times \{ r^{-1} [1 - (\alpha + \theta)r^{-1}]^{-1} \\
&\quad \quad \quad - r[1 - (\alpha + \theta)r]^{-1}(\alpha + \theta) \\
&\quad \quad \quad + (\alpha + \theta) \\
&\quad \quad \quad \times \{ r[1 - (\alpha + \theta)r]^{-1} \\
&\quad \quad \quad \quad - r^{-1} [1 - (\alpha + \theta)r^{-1}] r^{2T} \} \\
&\quad \quad \left. + 1 - r^T \right\} \Big).
\end{aligned}$$

Therefore,

$$\begin{aligned}
 (A8) \quad \lim_{T \rightarrow \infty} \hat{A} &= 2[\sigma_w^2/(1-r^2)] \\
 &\times [1-(\alpha+\theta)^2]^{-1} \\
 &\times \{2(\alpha+\theta)r \\
 &\times [1-(\alpha+\theta)r]^{-1} + 1\}.
 \end{aligned}$$

Finally,

$$\begin{aligned}
 (A9) \quad \text{Cov}(\Delta^T k_t, \Delta^T l_t) &= (1-\alpha) \\
 &\times \sum_{j=0}^{\infty} \text{Cov}(\Delta^T l_t, \Delta^T l_{t-j-1}) \\
 &\times (\alpha+\theta)^j \\
 &= (1-\alpha)\sigma_\varepsilon^2 \{T[1-(\alpha+\theta)] \\
 &\quad - [1-(\alpha+\theta)^T]\} / [1-(\alpha+\theta)]^2 \\
 &\quad \text{for } r=1.
 \end{aligned}$$

In general, for arbitrary r ,

$$\begin{aligned}
 \lim_{T \rightarrow \infty} [\text{Cov}(\Delta^T k_t, \Delta^T l_t)] &= \lim_{T \rightarrow \infty} \left\{ [(1-\alpha)\sigma_\varepsilon^2/(1-r^2)] \right. \\
 &\times \left\{ (r^T-1)^2 r [1-(\alpha+\theta)r]^{-1} \right. \\
 &\quad + (r^T-1)r^{T+1} \\
 &\quad \times \left\{ r^{-2} [1-(\alpha+\theta)r^{-1}]^{-1} \right. \\
 &\quad \quad - [1-(\alpha+\theta)r]^{-1} \} \\
 &\quad \left. - r^{2T-1} [1-(\alpha+\theta)r^{-1}]^{-1} \right. \\
 &\quad \left. + r [1-(\alpha+\theta)r]^{-1} \right\} \}.
 \end{aligned}$$

Taking the limit, for $r < 1$,

$$\begin{aligned}
 (A10) \quad \lim_{T \rightarrow \infty} [\text{Cov}(\Delta^T l_t, \Delta^T k_t)] &= [(1-\alpha)2\sigma_\varepsilon^2/(1-r^2)] \\
 &\times [1-(\alpha+\theta)r]^{-1}.
 \end{aligned}$$

Table A1 summarizes the results of this appendix that are relevant for the bias described in equation (16) of the text.

Combined cases: $r, \rho \rightarrow 1$, $T \rightarrow \infty$. We shall now use the expression in the second column, and we shall send r^2 and ρ^2 to unity at the same rate. The resulting expressions are then used in equations (16) and (17) of the text (the Landau symbol "O" refers to the order of the expression):

$$\begin{aligned}
 a_{kk} &= O\left(\frac{1}{1-\rho^2}\right) \left(\frac{2\sigma_\varepsilon^2}{1-(\alpha+\theta)^2}\right) \\
 &\times \left(\frac{2(\alpha+\theta)}{1-(\alpha+\theta)} + 1\right) \\
 &+ O\left(\frac{1}{1-r^2}\right) \left(\frac{2\sigma_w^2(1-\alpha)^2}{1-(\alpha+\theta)^2}\right) \\
 &\times \left(\frac{2(\alpha+\theta)}{1-(\alpha+\theta)} + 1\right) \\
 &= O\left(\frac{1}{1-\rho^2}\right) \left(\frac{2\sigma_\varepsilon^2}{1-(\alpha+\theta)^2}\right) \\
 &\times \left(\frac{1+(\alpha+\theta)}{1-(\alpha+\theta)}\right) \\
 &+ O\left(\frac{1}{1-r^2}\right) \left(\frac{2\sigma_w^2(1-\alpha)^2}{1-(\alpha+\theta)^2}\right) \\
 &\times \left(\frac{1+(\alpha+\theta)}{1-(\alpha+\theta)}\right) \\
 a_{kl} &= O\left(\frac{1}{1-r^2}\right) \left(\frac{2\sigma_w^2}{1-(\alpha+\theta)}\right) (1-\alpha)
 \end{aligned}$$

TABLE A1—EXPRESSIONS FOR THE a_{ij} FROM WHICH ONE MAY CALCULATE THE BIAS IN \hat{b}_k AND \hat{b}_l

Parameter	Case (a): $\lim_{r \rightarrow 1} \lim_{\rho \rightarrow 1} T$ finite	Case (b): $\rho, r < 1, \lim_{T \rightarrow \infty}$
a_{kk} :	$\frac{2\sigma_e^2(\alpha + \theta) \left(\frac{T}{1 - (\alpha + \theta)} - \frac{1 - (\alpha + \theta)^T}{[1 - (\alpha + \theta)]^2} \right) + T\sigma_e^2}{1 - (\alpha + \theta)^2}$ $+ (1 - \alpha)^2 \left[\frac{2\sigma_w^2(\alpha + \theta) \left(\frac{T}{1 - (\alpha + \theta)} - \frac{1 - (\alpha + \theta)^T}{[1 - (\alpha + \theta)]^2} \right) + T\sigma_w^2}{1 - (\alpha + \theta)^2} \right]$	$\left(\frac{2\sigma_e^2}{(1 - \rho^2)[1 - (\alpha + \theta)^2]} \right) \left(\frac{2(\alpha + \theta)\rho}{1 - (\alpha + \theta)\rho} + 1 \right)$ $+ \left(\frac{(1 - \alpha)^2 2\sigma_w^2}{(1 - r^2)[1 - (\alpha + \theta)^2]} \right) \left(\frac{2(\alpha + \theta)r}{1 - (\alpha + \theta)r} + 1 \right)$
a_{kl} :	$(1 - \alpha)\sigma_w^2 \left(\frac{T[1 - (\alpha + \theta)] - [1 - (\alpha + \theta)^T]}{[1 - (\alpha + \theta)]^2} \right)$	$2(1 - \alpha) \left(\frac{\sigma_w^2}{1 - r^2} \right) [1 - (\alpha + \theta)r]^{-1}$
a_{ll} :	$T\sigma_w^2$	$2\sigma_w^2/(1 - r^2)$
a_{ku} :	$\frac{\sigma_w^2 \{ T[1 - (\alpha + \theta)] - [1 - (\alpha + \theta)^T] \}}{[1 - (\alpha + \theta)]^2}$	$\frac{2\sigma_e^2 \rho}{(1 - \rho^2)[1 - \rho(\alpha + \theta)]}$

Note: To obtain the a_{ij} , the expressions in the first column of the table [case (a)] should be divided by T^2 ; the second column reports $\lim_{T \rightarrow \infty} T^2 a_{ij}$ for case (b).

$$a_{ll} = O\left(\frac{1}{1 - r^2}\right)(2\sigma_w^2)$$

$$a_{ku} = O\left(\frac{1}{1 - \rho^2}\right)\left(\frac{2\sigma_e^2}{1 - (\alpha + \theta)}\right).$$

Therefore, letting A_{ij} be the constant in the expression for a_{ij} ,

$$\begin{aligned} & \frac{a_{ll}a_{ku}}{a_{kk}a_{ll} - a_{kl}^2} \\ &= O\left(\frac{1}{1 - r^2}\right)(A_{ll}) O\left(\frac{1}{1 - \rho^2}\right)(A_{ku}) \\ & \times \left\{ \left[O\left(\frac{1}{1 - \rho^2}\right)(A_{kk}^1) \right. \right. \\ & \quad \left. \left. + O\left(\frac{1}{1 - r^2}\right)(A_{kk}^2) \right] \right. \\ & \times O\left(\frac{1}{1 - r^2}\right)(A_{ll}) \\ & \quad \left. - \left[O\left(\frac{1}{1 - r^2}\right) \right]^2 (A_{kl}^2) \right\}^{-1} \end{aligned}$$

where A_{kk}^1 and A_{kk}^2 are the first and second terms in the expression for a_{kk} . Now send $1 - \rho^2$ and $1 - r^2$ to 0 at the same rate, to get

$$\begin{aligned} & \frac{a_{ll}a_{ku}}{a_{kk}a_{ll} - a_{kl}^2} \\ & \rightarrow \frac{A_{ll}A_{ku}}{(A_{kk}^1 + A_{kk}^2)A_{ll} - A_{kl}^2} \\ &= \left(\frac{4\sigma_w^2\sigma_e^2}{1 - (\alpha + \theta)} \right) \\ & \times \left\{ \left[\left(\frac{2\sigma_e^2}{1 - (\alpha + \theta)^2} \right) \left(\frac{1 + (\alpha + \theta)}{1 - (\alpha + \theta)} \right) \right. \right. \\ & \quad \left. \left. + \left(\frac{2\sigma_w^2(1 - \alpha)^2}{1 - (\alpha + \theta)^2} \right) \left(\frac{1 + (\alpha + \theta)}{1 - (\alpha + \theta)} \right) \right] \right. \\ & \quad \left. \times 2\sigma_w^2 - \frac{4\sigma_w^4(1 - \alpha)^2}{[1 - (\alpha + \theta)]^2} \right\}^{-1}. \end{aligned}$$

Observing that $1 - (\alpha + \theta)^2 = [1 - (\alpha + \theta)](1 + \alpha + \theta)$ and making that substitution on the bottom line of the above expression leads to $[1 - (\alpha + \theta)]^2$ entering everywhere in the bottom of the denominator (i.e., the expression in large braces). Then, multiplying top and bottom by $[1 - (\alpha + \theta)]^2 / 4\sigma_w^2$ leaves us with

$$\frac{\sigma_\varepsilon^2 [1 - (\alpha + \theta)]}{\sigma_\varepsilon^2 + (1 - \alpha)\sigma_w^2 - (1 - \alpha)\sigma_w^2} = 1 - (\alpha + \theta).$$

Substituting this into (16) leads to (17). We now calculate the bias on \hat{b}_l :

$$\begin{aligned} & \frac{-a_{kl}a_{ku}}{a_{kk}a_{ll} - a_{kl}^2} \\ &= -O\left(\frac{1}{1 - r^2}\right)(A_{kl}) \\ & \quad \times O\left(\frac{1}{1 - \rho^2}\right)(A_{ku}) \\ & \quad \times \left\{ O\left(\frac{1}{1 - \rho^2}\right)(A_{kk}^1) \right. \\ & \quad \left. + O\left(\frac{1}{1 - r^2}\right)(A_{kk}^2) \right\} \\ & \quad \times O\left(\frac{1}{1 - r^2}\right)(A_{ll}) \\ & \quad - \left[O\left(\frac{1}{1 - r^2}\right) \right]^2 A_{kl}^2 \Big\}^{-1} \\ & \rightarrow \frac{-A_{kl}A_{ku}}{(A_{kk}^1 + A_{kk}^2)A_{ll} - A_{kl}^2} \\ &= \frac{-4\sigma_w^2\sigma_\varepsilon^2(1 - \alpha)}{[1 - (\alpha + \theta)]^2} \\ & \quad \times \left\{ \left[\left(\frac{2\sigma_\varepsilon^2}{1 - (\alpha + \theta)^2} \right) \left(\frac{1 + (\alpha + \theta)}{1 - (\alpha + \theta)} \right) \right] \right. \\ & \quad \left. + \left(\frac{2\sigma_w^2(1 - \alpha)^2}{1 - (\alpha + \theta)^2} \right) \left(\frac{1 + (\alpha + \theta)}{1 - (\alpha + \theta)} \right) \right\} \\ & \quad \times 2\sigma_w^2 - \frac{4\sigma_w^2(1 - \alpha)^2}{[1 - (\alpha + \theta)]^2} \Big\}^{-1}. \end{aligned}$$

We note that

$$\begin{aligned} & [1 - (\alpha + \theta)^2][1 - (\alpha + \theta)] \\ &= (1 + \alpha + \theta)[1 - (\alpha + \theta)]^2. \end{aligned}$$

Therefore, the above equals

$$\begin{aligned} & \frac{-4\sigma_w^2\sigma_\varepsilon^2(1 - \alpha)}{2\sigma_w^2[2\sigma_\varepsilon^2 + (1 - \alpha)^2 2\sigma_w^2] - 4\sigma_w^2(1 - \alpha)^2} \\ &= \frac{-\sigma_\varepsilon^2(1 - \alpha)}{\sigma_\varepsilon^2 + (1 - \alpha)^2\sigma_w^2 - \sigma_w^2(1 - \alpha)^2} \\ &= -(1 - \alpha). \end{aligned}$$

When substituted into (16) this leads to (18).

Appendix 3

Here, we briefly describe our analysis of the equation $K_{t+1} = sY_t + (1 - \delta)K_t$. This analysis led to the estimation reported in Tables 3 and 4. Under this hypothesis,

$$\begin{aligned} y_{t+1} &= z_{t+1} + (1 - \alpha)l_{t+1} \\ & \quad + (\alpha + \theta)\ln[sY_t + (1 - \delta)K_t] \\ &= z_{t+1} + (1 - \alpha)l_{t+1} \\ & \quad + (\alpha + \theta)\ln\{sY_t + (1 - \delta) \\ & \quad \times [sY_{t-1} + (1 - \delta)K_{t-1}]\} \\ &= z_{t+1} + (1 - \alpha)l_{t+1} \\ & \quad + (\alpha + \theta)\ln s \\ & \quad + (\alpha + \theta)\ln \left[\sum_{j=0}^{\infty} (1 - \delta)^j Y_{t-j} \right]. \end{aligned}$$

Therefore, the analogue of the equation in footnote 11 is

$$\begin{aligned}
 (A11) \quad y_{t+1} - \rho y_t &= \omega_t + (1 - \alpha)(l_{t+1} - \rho l_t) \\
 &+ (\alpha + \theta)(1 - \rho) \ln s \\
 &+ (\alpha + \theta) \left[\sum_{j=0}^{\infty} (1 - \delta)^j Y_{t-j} \right. \\
 &\quad \left. - \rho \ln \sum_{j=0}^{\infty} (1 - \delta)^j Y_{t-j-1} \right].
 \end{aligned}$$

The ω_t process was once again assumed to follow equation (4). The infinite sums in (A11) were truncated at $j = 20$. This was possible because only yearly data are used in Tables 3 and 4. Since at least about 20 years of data are needed to construct a reasonable approximation to the infinite sum of past Y 's, we could not use quarterly data, as these are available only for the postwar years (see footnote 11, however).

REFERENCES

- Baily, Martin Neal, "Comment," *NBER Macroeconomic Annual*, 1987, 1, 205-8.
- Barro, Robert, "The Persistence of Unemployment," *American Economic Review*, May 1988, 78, 32-7.
- Baumol, William and Wolff, Edward, "Productivity Growth, Convergence, and Welfare: Reply," *American Economic Review*, December 1988, 78, 1155-9.
- Benhabib, Jess and Laroque, Guy, "On Competitive Cycles in Production Economies," *Journal of Economic Theory*, June 1988, 45, 145-70.
- and Nishimura, Kazuo, "Stochastic Equilibrium Oscillations," *International Economic Review*, February 1989, 30, 85-102.
- Bernstein, Jeffrey and Nadiri, Ishaq, "Research and Development and Intraindustry Spillovers: An Empirical Implication of Dynamic Duality," *Review of Economic Studies*, April 1989, 56, 249-68.
- Blume, Lawrence and Easley, David, "Characterization of Optimal Plans for Stochastic Dynamic Programs," *Journal of Economic Theory*, April 1982, 28, 221-34.
- Campbell, John and Mankiw, Gregg, "Are Output Fluctuations Transitory?" *Quarterly Journal of Economics*, November 1987, 102, 857-80.
- Christiano, Lawrence, "Comment on Romer's 'Crazy Explanations of the Productivity Slowdown'," unpublished manuscript, Federal Reserve Bank of Minneapolis, 1987.
- Cochrane, John, "How Big is the Random Walk in GNP?" *Journal of Political Economy*, October 1988, 96, 893-920.
- DeLong, Bradford, "Productivity Growth, Convergence, and Welfare: Comment," *American Economic Review*, December 1988, 78, 1138-54.
- Griliches, Zvi, "Issues in Assessing the Contribution of Research and Development to Productivity Growth," *Bell Journal of Economics*, Spring 1979, 10, 92-116.
- , "Productivity Puzzles and R&D: Another Nonexplanation," *Journal of Economic Perspectives*, Fall 1988, 2, 9-21.
- Heston, Alan and Summers, Robert, "Improved International Comparisons of Real Product and Its Composition," *Review of Income and Wealth*, June 1984, 30, 207-26.
- Jaffe, Adam, "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value," *American Economic Review*, December 1986, 76, 984-1001.
- Jovanovic, Boyan and MacDonald, Glenn, "Competitive Diffusion," unpublished manuscript, New York University, October 1988.
- and Rob, Rafael, "The Growth and Diffusion of Knowledge," *Review of Economic Studies*, October 1989, 56, 569-82.
- Lach, Saul and Schankerman, Mark, "Dynamics of R&D and Investment in the Scientific Sector," *Journal of Political Economy*, August 1989, 97, 880-904.
- Maddison, Angus, *Phases of Capitalist Development*, New York: Oxford University Press, 1982.
- Mansfield, Edwin, Rapoport, John, Romeo, Anthony, Wagner, Samuel and Beardsley, George, "Social and Private Rates of Re-

- turn from Industrial Innovations," *Quarterly Journal of Economics*, May 1977, 91, 221-40.
- Nelson, Charles and Plosser, Charles, "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics*, September 1982, 10, 139-62.
- Pakes, Ariel and Schankerman, Mark, (1984a) "The Rate of Obsolescence of Patents, Research Gestation Lags and the Private Rate of Return to Research Resources," in Zvi Griliches, ed., *R&D, Patents and Productivity*, Chicago: University of Chicago Press, 1984, 73-88.
- _____ and _____, (1984b) "An Exploration into the Determinants of Research Intensity," in Zvi Griliches, ed., *R&D, Patents and Productivity*, Chicago: University of Chicago Press, 1984, 209-32.
- Prescott, Edward C., "Theory Ahead of Measurement in Business-Cycle Research," *Carnegie-Rochester Conference on Public Policy*, Autumn 1986, 25, 11-44.
- Quah, Danny, "International Patterns of Growth: Persistence in Cross-Country Disparities," unpublished manuscript, Massachusetts Institute of Technology, January 1990.
- Romer, Paul, "Crazy Explanations for the Productivity Slowdown," *NBER Macroeconomics Annual*, 1987, 1, 163-201.
- Scherer, Frederick M., "Inter-Industry Technology Flows and Productivity Growth," *Review of Economics and Statistics*, November 1982, 64, 627-34.
- Schmookler, Jacob, *Invention and Economic Growth*, Cambridge, MA: Harvard University Press, 1966.
- Shleifer, Andrei, "Implementation Cycles," *Journal of Political Economy*, December 1986, 94, 1163-90.
- Solow, Robert, "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics*, August 1957, 39, 312-20.
- Vernon, Raymond, "Technological Development: The Historical Experience," EDI Seminar Paper 39, The World Bank, Washington, DC, 1989.

Research Productivity Over the Life Cycle: Evidence for Academic Scientists

By SHARON G. LEVIN AND PAULA E. STEPHAN *

The relationship between age and the publishing productivity of Ph.D. scientists is analyzed using data from the Survey of Doctorate Recipients (National Research Council) and the Science Citation Index. The longitudinal nature of the data allows for the identification of pure aging effects. In five of the six areas studied, life-cycle aging effects are present. Only in particle physics, where scientists often speak of being on a "religious quest," is there indication that scientific productivity is not investment-motivated. Vintage effects are also considered. The expectation that the latest educated are the most productive is not generally supported by the data. (JEL 022, 821, 841, 851)

Research productivity over the life cycle has become an increasingly important topic to the American scientific community as the average age of scientists affiliated with institutions of higher learning has increased.¹ A popular belief held by scientists and the lay

public alike is that science is a young person's game. Karl F. Gauss was 18 when he developed least-squares, Charles R. Darwin was 29 when he developed the concept of natural selection, Albert Einstein was 26 when he formulated the theory of relativity, and Sir Isaac Newton was 24 when he began his work on universal gravitation, calculus, and the theory of colors (Stephen Cole, 1979).

This paper examines the research productivity of scientists over the life cycle. We develop a model of scientific productivity in which scientists engage in research not only for the present value of the stream of future financial rewards associated with research, but also for the current satisfaction that research provides the scientist. The model is estimated with a unique pooled cross-section longitudinal data base² created from the National Research Council's biennial 1973-1979 *Survey of Doctorate Recipients* (SDR) for scientists at Ph.D.-granting institutions trained in six subfields of physics and earth science.

*Department of Economics, University of Missouri-St. Louis, St. Louis, MO 63121; and Policy Research Program and Department of Economics, Georgia State University, Atlanta, GA 30303, respectively. This work was supported by funds from the Alfred P. Sloan Foundation (B1983-43), the Exxon Education Foundation, and the National Science Foundation (grant no. SRS 8306947). Additional resources were provided by the Graduate School and Gerontology Program of the University of Missouri-St. Louis and the Georgia State University Research Office and College of Business Administration Research Council. All opinions expressed are those of the authors and not the granting institutions. We thank George Boyce and Sue Henn of the National Research Council for their assistance in putting together the data base. In addition to the individuals mentioned in Appendix A, we also acknowledge the help of Michael Pogodzinski, Rubin Saposnik, Frank Stafford, Lester Taylor, and members of the Emory University-Georgia State University Seminar. Graduate research assistants on the project were Robert Eisenstadt, Stacy Kottman, Jan Luytjes, and Yan Liao. Finally, we acknowledge the helpful comments of three anonymous referees.

¹Between 1968 and 1978, the proportion of young doctorates (those who received a degree in the past seven years) among science and engineering full-time faculty declined from about 43 percent to 26 percent (National Research Council, 1979 p. 26). Between 1975 and 1985, the proportion of doctoral scientists under 35 years of age engaged in teaching declined from 24 percent to 9 percent, while the proportion over 55 years of age increased from 14 percent to 24 percent (National Science Foundation, 1988 p. 44).

²The data base resides with the Data Processing Unit of the Office of Scientific and Engineering Personnel at the National Research Council (NRC). Upon request, we will provide the documentation necessary for replication. Use of the data base by researchers for purposes other than replication will likely require obtaining permission from the National Academy of Sciences and the Institute of Scientific Information, given the terms of a lease agreement signed in 1983.

The results support an investment model of scientific productivity. In five of the six areas, life-cycle aging effects are present. Only in particle physics, where scientists are often portrayed as on a "religious quest," is there any indication that scientific productivity is not investment-motivated. Assuming that demand conditions do not change markedly, these results suggest that the American scientific community over the next 10 or 15 years may not be as productive as a younger community was in the 1960's and early 1970's.

Section I presents a conceptual model of scientific productivity, incorporating both the investment and consumption motives noted above. Section II sets forth the methodology and specification used to estimate the model. Section III presents the results, and the conclusions follow in Section IV.

I. A Conceptual Framework

Two hypotheses are commonly advanced for scientists engaging in research. One focuses on research as investment-motivated, arguing that scientists engage in research because of the future financial rewards associated with the activity; the other focuses on research as consumption-motivated, downplaying the importance of financial rewards and stressing instead the scientist's fascination with the research puzzle itself. "Research is in many ways a kind of game, a puzzle-solving operation in which the solution of the puzzle is its own reward" (Warren Hagstrom, 1965 p. 16). Although the investment motive implies a decline in research productivity over the career, given the finite time horizon (Arthur Diamond, 1984), the consumption motive does not. The model set forth below incorporates both the investment and consumption motives for research.

(i) The individual chooses to allocate time between two activities, research and nonresearch, such as teaching and consulting, which produces current income.

(ii) The objective of the scientist is to allocate time in such a way as to maximize utility, U , over a career which begins at time zero, after receipt of the Ph.D., and

ends at time T , the day of retirement. Utility is a function of research output, R_t , and market goods, X_t , which cost a constant price, p . Thus, the problem is to choose s_t , the proportion of time engaged in producing research, so as to maximize

$$(1) \quad J = \int_0^T e^{-\rho t} U(R_t, X_t) dt$$

$$= \int_0^T e^{-\rho t} \ln(R_t^{\Theta_1} X_t^{\Theta_2}) dt$$

$$\Theta_1, \Theta_2 > 0.$$

We assume that T is known and, following Harl Ryder et al. (1976), also that ρ , the time-preference parameter, is zero. This particular utility function has been widely used in demand analysis (e.g., Robert Pollak and Terrence Wales, 1969; Ryder et al., 1976). Further, because of the high acceptance rates in science (Carnot Nelson and Dennis Pollock, 1970; A. Carolyn Miller and Sharon Serzan, 1984), we make the simplifying assumption that all research output is published.

(iii) Although publications do not wear out, their relevance does, as change occurs in the field. Thus, changes in the stock of publications deemed relevant, P_t , is given by

$$(2) \quad \dot{P}_t = R_t - \delta P_t$$

where δ is a depreciation rate and the dot denotes the derivative with respect to time.

(iv) Income in any time period is a function of P_t . Previous publications that are no longer valued by the field do not contribute to current income, I_t , where $I_t = \alpha(1 - s_t)P_t$ and α is the rental value of a unit of P .

(v) The change in assets, A , over time, is given by

$$(3) \quad \dot{A}_t = -pX_t + \alpha(1 - s_t)P_t + rA_t$$

where r is the interest rate.

(vi) A by-product of producing research is learning. Because most research has non-human-capital-enhancing aspects, only h

proportion of the research output that is produced at time t is incorporated into the individual's stock of knowledge. As a result, K_t , the scientist's effective knowledge (knowledge that is "up to date") equals hP_t .

(vii) Research output is produced by combining effective knowledge with time

$$(4) \quad R_t = f(s_t K_t) \\ = A_1 (s_t h P_t)^\beta \quad \beta < 1, A_1 > 0.$$

This production function has a long tradition in human capital studies (e.g., Yoram Ben-Porath, 1967; Ryder et al., 1976; Stephan, 1976; John McDowell, 1982; Diamond, 1984).

With the problem as formulated in (i)–(vii), a dynamic model is necessary. In terms of control theory, the Hamiltonian takes the form

$$(5) \quad H = \ln R_t^{\Theta_1} + \ln X_t^{\Theta_2} \\ + \lambda_{P_t} (A_1 s_t^\beta h^\beta P_t^\beta - \delta P_t) \\ + \lambda_{A_t} [-pX_t + \alpha(1-s_t)P_t + rA_t]$$

where λ_{A_t} and λ_{P_t} represent the shadow values of assets and effective publications, respectively. The necessary conditions for a maximum are to let \hat{s}_t and \hat{X}_t be control variables that maximize the objective function, subject to the conditions given by (2), (3), and (4), some initial stock of articles (coauthored, as is the custom in science, with advisors while in graduate school), and some initial stock of assets.³ Then, \hat{s}_t and \hat{X}_t maximize the Hamiltonian, and the shadow prices satisfy the equations

$$(6) \quad \dot{\lambda}_{P_t} = -\partial H / \partial P_t = -\beta \Theta_1 / P_t \\ - (\lambda_{P_t} \beta R_t) / P_t + \delta \lambda_{P_t} - \lambda_{A_t} \alpha (1-s_t)$$

$$(7) \quad \dot{\lambda}_{A_t} = -\partial H / \partial A_t = -r.$$

³There are, of course, two stages to this problem. The first stage is one of complete specialization in research and is characterized by $s_t = 1$. In the second stage, the time constraint is not binding. Here, we focus on the second stage and assume that the scientist stays in graduate school until ready to pass from stage 1 to stage 2. If, however, $\Theta_2 = 0$, the scientist never is motivated to earn income, and stage 1 persists for the duration of the career.

The optimal path is the solution to the differential equations in A , P , λ_P , and λ_A satisfying the transversality conditions that

$$(8) \quad \lambda_{P_T} P_T = 0 \quad (P_T = \text{free}, \lambda_{P_T} = 0)$$

and

$$(9) \quad \lambda_{A_T} A_T = 0 \quad (A_T = 0, \lambda_{A_T} = \text{free}).$$

Assuming that $r = 0$ in order to simplify the problem, it follows that λ_{A_t} is constant over time.

Maximizing the Hamiltonian with respect to X yields the result that market goods consumed by the individual are constant over time:

$$(10) \quad \hat{X}_t = \Theta_2 / \lambda_{A_t} p$$

where the caret denotes the optimal value. Maximizing the Hamiltonian with respect to s and substituting the value of s into (6) yields

$$(11) \quad \dot{\lambda}_{P_t} = \delta \lambda_{P_t} - \lambda_{A_t} \alpha.$$

Given the transversality conditions, this implies that

$$(12) \quad \lambda_{P_t} = (\lambda_{A_t} \alpha / \delta) [1 - e^{-\delta(T-t)}]$$

and

$$(13) \quad \hat{R}_t = A_1 [\Theta_1 E + (h\beta / \delta) \\ \times \hat{R}_t (1 - e^{-\delta(T-t)})]^\beta$$

where $E = [h\beta / (\lambda_{A_t} \alpha)] > 0$. Thus, in this model, $\hat{R}_t > 0$, even at time T , the date of retirement, when $\hat{R}_T = A_1 [\Theta_1 E]^\beta$.

Although there is no known solution for equation (13), two propositions can be derived.

PROPOSITION 1: *Research activity is greater, the greater is the taste for research, Θ_1 :*

$$(14) \quad \frac{\partial \hat{R}_t}{\partial \Theta_1} = \frac{\beta \hat{R}_t E}{\Theta_1 E + X} > 0 \quad \text{for } t < T$$

where $X = (h\beta/\delta)R_t(1 - \beta)(1 - e^{-\delta(T-t)}) > 0$ for all $t < T$ and $X = 0$ for $t = T$.

The sociology-of-science literature makes the argument that this taste for research is learned in graduate school. Moreover, interviews with scientists suggest that this taste is field-dependent, with scientists in some fields placing greater value on the satisfaction derived from engaging in research than do scientists in other fields.

PROPOSITION 2: *Research activity declines over the life cycle:*

$$(15) \quad \dot{R}_t = \frac{(-\beta^2 \hat{R}^2 h) e^{-\delta(T-t)}}{\Theta_1 E + X} < 0$$

for $t < T$.

Furthermore, although it is not possible to prove, simulations of the model support the concept that larger values of Θ_1 , the taste for research, generate flatter research profiles. This has some intuitive appeal. Early in the career, the strong investment incentive for research complements a scientist's puzzle-solving urge, but as the scientist ages and the present value of the investment benefit declines, the scientist must supplement the investment component to meet the puzzle-solving need. Clearly, this is more important the larger is Θ_1 .⁴

⁴Two special cases in which a utility parameter is zero are also of interest. First, when $\Theta_1 = 0$ and the scientist derives no satisfaction from engaging in research, the path of R_t is the same as the path in a simplified Ben-Porath income-maximizing model. In this case, $R_T = 0$, and the simulations suggest that the R_t profile is steeper than when $\Theta_1 > 0$. Second, when $\Theta_2 = 0$ and market goods provide no satisfaction, the scientist lacks the motivation to earn income and remains in the stage of complete specialization throughout the career (see footnote 3). In this situation, $s_t = 1$, and research is constrained only by the amount of effective knowledge available to the scientist. Thus,

$$R_t = A_1(hP_t)^\beta \quad \text{and} \quad \dot{R}_t = \beta R_t/P_t[R_t - \delta P_t].$$

It follows in this case that, as long as $R_t > \delta P_t$, research increases with career age. Simulations suggest that $R_t > \delta P_t$ for a period longer than a normal career for reasonable values of P_0 , A_1 , β , and h .

II. Methodology

Research productivity over the life cycle has received little attention in the economics literature, although there have been several related studies in other disciplines (e.g., Harvey Lehman, 1953; Paul Allison and John Stewart, 1974; Alan Bayer and Jeffrey Dutton, 1977; Harriet Zuckerman, 1977; Stephen Cole, 1979; Allison et al., 1982; Diamond, 1986). To date, the empirical evidence on the life-cycle effect is weak and largely inconclusive, because most studies use cross-sectional data. Since scientists of different ages come from different cohorts in a cross-sectional study, aging effects are confounded with cohort effects.

One type of cohort effect is associated with change in the knowledge base of the scientist's field. Because of what Jacob Mincer (1974 p. 21) calls the "secular progress of knowledge," there is a general presumption in science that the latest educated are the best educated. We incorporate this concept by making the depreciation rate in the model (δ) vintage-dependent, so that successively later cohorts face lower and lower depreciation rates as science progresses toward a more complete understanding of the "laws" governing the universe.⁵ Thus, the stock of relevant articles (2) declines more rapidly for scientists coming from earlier vintages than for scientists coming from later vintages. Since

$$(16) \quad \frac{\partial \hat{R}_t}{\partial \delta} = (\beta \hat{R}_t / \delta)^2 h \times \frac{(e^{-\delta(T-t)})(\delta(T-t) + 1) - 1}{\Theta_1 E + X} < 0$$

for all $t < T$, it follows that $\partial \hat{R}_t / \partial V > 0$, where V stands for the date of the Ph.D.⁶

⁵Clearly, this is a simplified view of how change in knowledge affects a scientist's stock of relevant publications. One might expect, for example, that the earliest articles authored by a scientist would be the first to be rendered obsolete.

⁶Previous work in human capital (Solomon Polachek, 1976; McDowell, 1982; Stephan and Levin, 1983) has examined variation in the rate at which knowledge becomes obsolete across academic fields or

Another factor that affects research productivity and varies by cohort is the state of the job market at the time the doctorate is received. The link in this instance between productivity and cohort is the strong evidence that research output is affected not only by attributes of the scientists but also by attributes of the employing institutions (J. Scott Long, 1978; Gerald Cole, 1979; Long and Robert McGinnis, 1981). Consequently, scientists graduating when appointments in the top academic sector are few and far between are expected to be less productive over their lives than are scientists who have the good fortune to leave graduate school when the prospects for employment in the top sector are good.

Finally, in addition to differences in the rate at which knowledge becomes obsolete and differences in opportunities that greet different cohorts over time, cohorts may vary in the levels of ability or motivation they bring to the fields or specialty areas they enter. Whether or not this occurs depends in part upon the desirability of other fields or professions relative to the one in question at the time the career decision is made. In recent years, this may have become a significant factor in science given the surge of interest in the high-salaried professions of business, law, and applied science. As a consequence, it is possible that these areas were able to secure a disproportionate share of the best minds from later cohorts, leaving some areas of science with a relative "brain drain."⁷

occupations and the consequences of this variation for investment over the life cycle as well as for career choice. While recognizing that rates of obsolescence vary across disciplines, here we focus on the idea that knowledge acquired during graduate school is more durable for later vintages than for earlier vintages, based on the assumption that science progresses towards "truth." Clearly, other issues may fall under the rubric of obsolescence (e.g., the distinction between anticipated obsolescence and unanticipated obsolescence and whether unanticipated change impacts relatively more on younger than on older scientists). Given our focus on aging, such issues are outside the scope of this article.

⁷See, for example, Howard Bowen and Jack Schuster (1986). In addition, in a recent interview on National Public Radio, Leon Lederman, a 1988 physics Nobel laureate, expressed concern over precisely this issue.

One way to control for vintage and these other cohort-related effects is to follow a single cohort over time. However, this approach ignores the fact that the scientific "state of the arts" and the work environment change over time. Thus, these calendar-time effects may also obscure the relationship between research productivity and age.

In this study, we develop a pooled cross-section time-series data base which permits us to control for cohort as well as for calendar time. The data base was created by matching records from the National Research Council's biennial 1973–1979 *Survey of Doctorate Recipients* (SDR) with publishing information from the *Science Citation Index* (SCI) prepared by the Institute of Scientific Information (ISI).⁸ (Details are provided in Appendix A). Four measures of research output were created. PUB1 measures straight publication counts occurring over a two-year period. PUB2 adjusts these counts for coauthorship. PUB3 uses Eugene Garfield's (1976) impact factor (ISI) to adjust for journal quality, while PUB4 adjusts the straight counts for both coauthorship and quality.

Although the data base was initially assembled for scientists trained in biochemistry, earth science, physics, and plant and animal physiology, the econometric investigation focused on six areas: solid-state/condensed-matter physics, particle physics, atomic and molecular physics, oceanography, geophysics, and geology.⁹ Subfield analyses were conducted because publishing patterns vary significantly across fields and subfields and the identification of vintage effects is feasible only at the subfield or specialty level. For each subfield, a case study identified vintage as well as other cohort effects (see Appendix A).

⁸Although the SDR is the largest and most comprehensive longitudinal study of scientists in the United States, it could not be used in previous studies of scientific productivity, because it did not contain measures of scientific productivity.

⁹Resource constraints and issues of confidentiality prevented us from studying other subfields or including women in the study.

The empirical analysis focuses on the research productivity of scientists employed full-time at prestigious doctoral-granting departments in their fields, for it is within this sector that the vast majority of research, at least in terms of journal publications, is produced. Thus, it was necessary to make a correction for nonrandom sample selection (James Heckman, 1979; Randall Olsen, 1980). Without this correction, life-cycle aging effects would be biased upward toward zero, since age and ability (which cannot be measured) are likely to be positively correlated in the selected sample, given that elite universities tend to hire relatively many young professors but retain only the best. Before pooling the data from each survey, we correct for this bias by estimating for each of the survey years an ordinary least-squares (OLS) regression predicting the likelihood that a scientist is employed in the selected sector and calculating Olsen's (1980) selectivity-bias correction variable. Included in each regression, among the other variables discussed in Appendix B, are categorical variables to capture the differences in job-market conditions experienced by Ph.D. cohorts.¹⁰

The model we wish to estimate takes the general form

$$(17) \quad R_{it} = f(\text{AGE}, V, T, X, S, u)$$

where R is a measure of publishing productivity of scientist i at time t , V is vintage, T is the calendar year, X is a vector of other explanatory variables suggested by the conceptual model, S is the sample-selection correction variable, and u is a stochastic

error term. The estimation strategy chosen addresses three issues: the identification of aging effects given that age, vintage, and calendar time are nearly perfectly correlated,¹¹ the presence of a limited dependent variable, and the presence of unmeasurable individual-specific fixed effects.

First, if scientists enter the labor force at the same age (APhD) and work continuously, then the calendar year (T) equals age minus APhD plus the year of doctorate, V . Because of this linear dependence ($T = \text{AGE} - \text{APhD} + V$), the effects of age, vintage, and calendar time on research productivity cannot be identified separately (William Johnson, 1980). One solution "is to assume non-linearity in the vintage effects" (Johnson, 1980 p. 401). Alternatively, identification is possible if at least one of the three variables "can be eliminated in favor of the underlying theoretical concepts" (Willard Rodgers, 1982 p. 783). Here the two solutions are merged. Following Johnson (1980), Ph.D. cohorts are grouped into intervals represented by categorical variables. The vintage effects are then modeled as a step function with each step corresponding to an interval. Rather than use arbitrary, five-year intervals as does Johnson, however, we used the case studies to identify Ph.D. classes which received a relatively homogeneous knowledge base in graduate school and thus shared a common likelihood of experiencing knowledge-obsolescence when change occurred in their field.¹²

¹¹The three effects are not perfectly correlated here, since scientists do not obtain the doctorate and enter the labor force at the same age.

¹²This required ascertaining when major changes in theory or technique occurred in the knowledge base of each subfield. Thus, if the case study for the subfield suggested that a major change occurred in 1949, another in 1955, and the last in 1967, we would group scientists in this field into four intervals and construct corresponding categorical variables: V_1 set equal to 1 if the degree was awarded after 1967, 0 otherwise; V_2 set equal to 1 if the degree was awarded between 1956 and 1967, 0 otherwise; V_3 set equal to 1 if the degree was awarded between 1950 and 1955, 0 otherwise. V_4 , the excluded (comparison) group would include those scientists educated prior to 1950. In terms of the latest-educated-are-best-educated framework, the step function should rise with each successively later vintage.

¹⁰To make this model tractable (see discussion on multiple criteria for selectivity in G. S. Maddala [1983 pp. 278–83]), we make the following assumptions. First, Ph.D. scientists desire the best academic jobs (see, e.g., R. A. Alpher et al., 1979; Beverly Porter, 1979a,b), and within academia, for the most part, mobility is downwards from the more prestigious to the less prestigious institutions. Thus, to be in the selected sample, individuals must have been chosen by these elite institutions and must have met the unwritten standards for continued employment, standards which are expected to tighten or loosen according to the state of the academic job market. Additional details concerning estimation are presented in Appendix B.

A second issue is the presence of a limited dependent variable, since some scientists in the SDR do not publish, or, at least, in certain years do not publish. Consequently, research productivity is truncated at zero, and OLS estimation would result in biased and inconsistent parameter estimates. Thus, we use the maximum-likelihood Tobit (James Tobin, 1958) procedure to estimate the parameters of the research-productivity model.

Finally, individual-specific unmeasurable effects are also an issue. It is well known that some scientists are extremely productive, while others are not. One reason for this is that some scientists possess a particular talent for research, a unique combination of creativity and motivation, which others do not possess (Mary Fox, 1983). Although this special talent may be randomly distributed and uncorrelated with age and vintage in the population of all scientists, this is unlikely in the censored samples considered here.¹³ Thus, we chose a fixed-effects estimator, rather than a random-effects (variance-components) estimator, because we expect that the unmeasurable individual-specific effects are correlated with other determinants of publishing productivity, and as a result, the variance-components estimator would be biased and inconsistent (George Judge et al., 1980).

Conceptually, this model can be estimated by including $N-1$ dummies for the individual-specific effects and $T-1$ dummies for the calendar-time effects in a Tobit specification. Inclusion of these fixed effects, however, complicates the analysis in three ways. First is the necessity that there be at least two observations per scientist in order to include the individual-specific dummies.¹⁴ Thus, in assembling the field

samples, only those scientists who were in the selected sector more than once were included.¹⁵ We refer to this as sample one in the findings reported below. A second issue that arises in the Tobit specification is that it is not possible to estimate an individual-specific fixed effect for a scientist who never published over the period surveyed. Consequently, a few additional cases were dropped from the analysis, and sample two was formed. (Note, however, that sample two still contains scientists who did not publish in some periods; thus, there remain numerous cases in which the dependent variable is truncated at zero.) Finally, because vintage itself is an individual-specific fixed effect, it proved impossible to obtain an estimate of its effect separate from the other unmeasurable individual-specific fixed effects in the model.¹⁶

Because of these complications, we chose to estimate two models using Tobit. Model A estimates the life-cycle publishing-productivity relationship for scientists in the elite sector and includes vintage dummy

ing-productivity model in each field does not differ for the two groups of scientists, the excluded and included groups, does not lead to a rejection of the null hypothesis.

¹³Note that we do not have a balanced design; scientists may be in the selected sample two, three, or four times. Although an unbalanced design necessitates an adjustment in estimating variance components (Mark Bils, 1985 p. 685), no adjustment is needed in the fixed-effects specification used here.

¹⁴Heckman and Thomas Macurdy (1980 p. 56) suggest that it would be possible to retrieve separate estimates for the vintage effects indirectly, after controlling for all the individual fixed effects, by regressing the estimated fixed effects obtained in model B on vintage. Our attempt to do so failed, however, because we have too few observations per individual. It also is not possible to obtain separate estimates for the vintage effects directly while controlling for the individual-specific effects in model B by dropping the time-period dummies. Since the vintage of the scientist does not change over time, no vintage estimate can be obtained when individual dummies for the fixed effects are included in the model. Finally, we note that Jerry Hausman and William Taylor (1981) suggest an alternative instrumental-variables technique by which it might be possible to estimate both the time-varying and time-invariant determinants of publishing productivity while controlling for the individual fixed effects. Given the time limitations of our data and our focus on aging, not vintage, we did not use their approach.

¹³For example, this latent, unobservable variable may be correlated with such determinants of publishing productivity as whether the scientist is employed at a Ph.D.-granting institution (i.e., the sample selection variable) or whether the scientist has garnered research support or reputational and positional prestige.

¹⁴Excluding cases of scientists who were in the selected sector only once raises again the possibility of sample selectivity bias. As discussed in Appendix B, a test of the conjecture that the structure of the publish-

variables in addition to the calendar-time dummies and the model's other parameters: age (AGE) and proxies (discussed in Appendix B) for the scientist's research environment (REPRANK), research effort (TEACH/ADMIN), research support (FEDSUP), and previous productivity (SALARY):¹⁷

$$(18) \quad R_{it} = c_1 + c_2 \text{AGE} + c_3 \text{REPRANK} \\ + c_4 \text{TEACH/ADMIN} \\ + c_5 \text{FEDSUP} + c_6 \text{SALARY} \\ + c_7 S_{it} + c_8 T_2 + c_9 T_3 \\ + c_{10} T_4 + c_g V_g + e_{it}$$

where T_2 , T_3 , and T_4 are the calendar-time dummies and V_g ($g = 1, 2, \dots, G$) are categorical variables from each case study, denoting different vintages of human capital. Since we did not control for the unmeasurable individual-specific fixed effects in this model, the inferences drawn about vintage effects must be viewed with caution. Model A is estimated using both samples one and two. (Because sample two is slightly smaller, in some cases the vintage categories had to be recombined.) The resulting estimates are referred to as A-1 and A-2 below.

Model B provides for a consistent estimate of the pure aging effect by including dummy variables to control for differences in the mean level of publishing productivity attributable to unmeasurables such as talent or motivation.¹⁸ This is done by dropping

the vintage variables from model A and including instead the individual-specific dummy variables, D_{i-1} , which capture all possible fixed individual-specific effects:¹⁹

$$(19) \quad R_{it} = c_1 + c_2 \text{AGE} + c_3 \text{REPRANK} \\ + c_4 \text{TEACH/ADMIN} \\ + c_5 \text{FEDSUP} + c_6 \text{SALARY} \\ + c_7 S_{it} + c_8 T_2 + c_9 T_3 \\ + c_{10} T_4 + c_{i-1} D_{i-1} + e_{it}$$

By comparing the restricted model A estimated using sample two (A-2) with the less-restricted model B, also estimated using sample two, we can test for the statistical significance of these additional, individual fixed effects and the robustness of the life-cycle effects observed in model A.

III. Findings

Space precludes the presentation of the econometric findings for all four output measures and for all six subfields. Instead, we illustrate our results focusing only on the parameter estimates for age, vintage, and calendar-time effects for one output measure, PUB1, the two-year count of journal publications.²⁰ The complete findings for all

would be likely to reflect differences in productivity between older and younger scientists, rather than differences in productivity as the average scientist aged (i.e., the pure aging effect), since the "between" variation in publishing productivity is much larger than the "within" variation.

¹⁹In estimating model B, care had to be taken to avoid singularities in the data matrix because of the multiple categorical variables representing time period and individual-specific fixed effects. Thus, in some cases, the time-period categories were collapsed, and as a result, the included categorical variables must be interpreted relative to a new omitted category.

²⁰These findings control for the likelihood of non-random sample selection discussed earlier. Although the sample-selection control variable was not always statistically significant, we found supportive evidence of bias for some form of the model estimated in 24 of the 28 possible combinations of subfields (particle physics was split into two groups; see discussion of findings) and output measures.

¹⁷Note that SALARY is, in effect, a lagged variable in the estimating equation, since research productivity is measured beginning one year after the survey date. However, one cannot automatically assume the exogeneity of SALARY. Applying the Hausman (1978) specification test in model A-1, however, we found that the null hypothesis of exogeneity could not be rejected for any output measure in any field. (The quality of graduate training and the age at time of Ph.D. were used as alternative instruments for SALARY.)

¹⁸Strictly speaking, consistency is a property of large samples and would require a large T , which is not possible in most empirical work (Maddala, 1987). If one failed to control for these individual fixed effects, we suspect that the age coefficient in the pooled model

TABLE 1—SELECTED REGRESSORS EXPLAINING PUBLICATION COUNTS (PUB1) FOR SUBFIELDS OF PHYSICS

Subfield and model	AGE	AGE × AGE	T_2	T_3	T_4	$\frac{V_1}{VIN1}$	$\frac{V_2}{VIN2}$	V_3	V_4	N	$\log L$	LR test (r, X^2)
Solid-state physicists:												
A-1 (V_1-V_4)	-0.399 ^c (0.102)		2.277 ^c (0.849)	2.861 ^c (0.978)	4.271 ^c (1.293)	-5.545 (4.145)	6.381 ^a (3.451)	7.055 ^b (3.010)	7.818 ^c (2.814)	182	-368.444 ^c	
A-2 (VIN1, VIN2)	-0.434 ^c (0.105)		2.014 ^b (0.900)	2.632 ^c (0.993)	4.172 ^c (1.326)	-0.962 (1.192)	0.224 (1.383)			159	-328.904 ^c	
B	2.431 ^c (0.735)	-0.027 ^c (0.009)		-0.091 (0.869)	-1.822 (1.719)					159	-237.322 ^c	(51, 174.789) ^c
Particle physicists at Ph.D.-granting institutions:												
A-1	-0.324 ^c (0.091)		0.934 (0.692)	2.515 ^c (0.822)	-2.878 (1.790)	-2.687 (2.200)	-0.574 (1.604)	0.214 (1.141)		168	-306.670 ^c	
A-2	-0.291 ^c (0.091)		0.822 (0.729)	1.752 ^c (0.802)	3.118 ^c (1.181)	-2.368 (2.229)	0.329 (1.632)	0.699 (1.164)		149	-303.371 ^c	
B	0.025 (0.279)		0.318 (0.870)	1.110 (1.418)						149	-229.442 ^c	(51, 97.360) ^c
Particle physicists at FFRDC's:												
A-1	-0.499 ^c (0.164)		1.117 (1.225)	2.367 ^c (1.388)	-0.594 (2.883)	-8.700 ^b (4.169)	-8.217 ^b (3.254)	-5.851 ^c (2.183)		157	-289.160 ^c	
A-2	-0.494 ^c (0.170)		0.748 (1.239)	2.054 (1.354)	0.307 (1.715)	-9.703 ^b (4.173)	-9.326 ^c (3.316)	-4.017 ^a (2.265)		117	-251.148	
B	-0.839 ^c (0.334)			2.421 ^b (1.096)	3.337 ^a (1.879)					117	-174.891 ^c	(39, 152.514) ^c
Atomic and molecular physicists:												
A-1	-0.164 ^a (0.117)		1.249 (1.460)	2.127 (1.501)	4.188 ^b (1.875)	1.264 (1.965)				89	-172.207 ^c	
A-2	-0.060 (0.131)		1.069 (1.493)	1.290 (1.520)	2.735 (1.905)	1.661 (1.953)				77	-163.590 ^c	
B	1.339 ^a (0.906)	-0.017 ^a (0.011)		-0.251 (1.283)	0.050 (2.067)					77	-114.889 ^c	(22, 95.358) ^c

Notes: Variable definitions and descriptive statistics are found in Appendix B (Table B1). The likelihood-ratio test (LR test) reports the number of restrictions (r) and the chi-square statistic for the comparison of models A-2 and B. All tests of significance are one-tailed, with the exception of time period and vintage effects. Standard errors are in parentheses.

^aStatistical significance at 0.10; ^bat 0.05; ^cat 0.01.

TABLE 2—SELECTED REGRESSORS EXPLAINING PUBLICATION COUNTS (PUB1) FOR SUBFIELDS OF EARTH SCIENCE

Subfield and model	AGE	AGE × AGE	T_2	T_3	T_4	$\frac{V_1}{VIN1}$	$\frac{V_2}{VIN2}$	$\frac{V_3}{VIN3}$	V_4	V_5	N	$\log L$	LR test (r, X^2)
Oceanographers:													
A-1 (V_1, V_2)	-0.067 ^a (0.049)		-1.270 (1.181)	-0.074 (0.784)	-0.687 (0.892)	-3.125 ^c (1.181)	-2.310 ^b (1.005)				57	-83.466	
A-2 (VIN1)	-0.002 (0.054)		-0.860 (0.881)	0.085 (0.835)	-1.155 (0.902)	0.185 (0.728)					51	-73.492	
B	0.928 ^b (0.553)	-0.020 ^c (0.008)		2.808 ^c (0.910)	3.194 ^b (1.441)						51	-48.945 ^c	(11, 40.251) ^c
Geophysicists:													
A-1 (V_1-V_4)	-0.461 ^c (0.128)		4.287 ^c (1.382)	3.158 ^b (1.472)	4.487 ^c (1.582)	-5.393 (4.475)	-2.425 (3.637)	-2.223 (2.764)	-2.146 (2.828)		78	-151.629 ^c	
A-2 (VIN1-VIN3)	-0.322 ^c (0.120)		4.171 ^c (1.335)	2.954 ^b (1.424)	4.134 (1.538)	-2.049 (3.659)	0.069 (2.803)	-0.325 (2.008)			69	-146.077 ^c	
B	2.370 ^c (0.779)	-0.020 ^b (0.009)		-2.139 ^b (1.052)	-2.644 (1.660)						69	-95.747 ^c	(18, 98.097) ^c
Geologists:													
A-1	-0.081 (0.075)		-0.874 (0.654)	0.494 (0.691)	0.524 (0.820)	2.013 (2.705)	2.675 (2.396)	1.527 (2.025)	1.816 (1.801)	1.702 (1.349)	172	-231.664 ^c	
A-2	0.097 (0.097)		-1.063 (0.682)	-0.281 (0.739)	-0.650 (0.896)	6.224 ^b (3.129)	6.136 ^b (2.820)	4.389 ^a (2.329)	3.966 ^a (2.025)	2.700 ^a (1.431)	130	-204.921 ^b	
B	-0.383 ^a (0.267)			1.651 ^a (0.969)	1.902 (1.460)						130	-172.085 ^c	(33, 65.671) ^c

Notes: Variable definitions and descriptive statistics are found in Appendix B (Table B2). The likelihood-ratio test (LR test) reports the number of restrictions (r) and the chi-square statistic for the comparison of models A-2 and B. All tests of significance are one-tailed, with the exception of time period and vintage effects. Standard errors are in parentheses.

^aStatistical significance at 0.10; ^bat 0.05; ^cat 0.01.

output measures are available upon request. Although the general conclusions are not particularly sensitive to the output measure used, differences for the other output measures are noted. Table 1 summarizes the findings for the physics areas investigated, and Table 2 summarizes the earth science results. Appendix B and Tables B1 and B2 describe the variables and report descriptive statistics for each of the fields.²¹

A. Physics

Three areas in physics are investigated: solid-state/condensed-matter physics, particle physics, and atomic and molecular physics. In layman's terms, solid-state/condensed-matter physics studies why substances have certain electrical properties, as well as other properties such as color and translucence. It is the largest subfield in physics, and research in this area is responsible for the transistor and superconductors, two of the most commercially viable developments in physics. Elementary particle physics focuses on the smallest bits of matter that are known to exist. Research in elementary particle physics looks for the laws governing the four fundamental interactions—nuclear (strong), electromagnetic, weak, and gravitational—with the final aim of unifying these interactions by finding some common origin. Abstract theorists working on unification are often depicted as involved in a "religious quest," handed them by Einstein, or, as is commonly stated in the literature, the "search for the Holy Grail." The fundamental equations which concern atomic and molecular physicists come from quantum mechanics. As a result, the equations have been known for approximately 60 years, although the solutions remain elusive. Theoretical atomic physicists continue to seek solutions to these equations.

Solid-State/Condensed-Matter Physics. Overall, the results are strong. The null hypothesis that the parameters of the model are jointly zero can be rejected at a confi-

dence level exceeding 0.99. The coefficients on the time variables in A-1 and A-2 indicate that output has increased in each successive time period (by about 2–4 articles) compared to the earliest period, 1973. More-recent vintages are more productive (by about 6–8 articles) than the earliest vintage (the omitted category), those educated prior to 1948, although the difference is only statistically significant for all measures of output for V_3 and V_4 . This is consistent with the case study's conjecture that the introduction of new experimental techniques as well as many-body theory and renormalization may have had the effect of depressing the output of the pre-1948 vintage. There is, however, little indication that publishing productivity varies significantly among later vintages. Not surprisingly, when the vintages are compressed into just three categories in A-2, there is no statistical evidence of vintage effects. The results also suggest the presence of other individual fixed effects in addition to the specified vintage effects. The likelihood-ratio test comparing model A-2 and model B indicates that the null hypothesis that there are no individual-specific unmeasurable fixed effects, after controlling for vintage, can be rejected at a level of significance exceeding 0.01.

Of particular interest to this study is the coefficient on aging. In A-1 and A-2, there is strong evidence of life-cycle effects, a decline of 0.4 articles per period. When the fixed-effects model is estimated (model B), the life-cycle effects persist. The coefficients on age and age-squared suggest a nonlinear aging effect, with publishing productivity reaching a peak at age 45. (For PUB2, the peak comes at 41, for PUB3 at 45, and for PUB4 at 40.)

Particle Physics. The prestigious sector in this area consists of scientists employed both at Ph.D.-granting institutions and at federally funded research and development centers (FFRDC's). These two groups are studied separately, since their research environments differ considerably.²²

²¹A full description of the subfield samples and descriptive statistics for the other explanatory variables are available upon request.

²²We did not consider and, in fact, had no way of modeling the possibility of endogenous sector choice in

(i) *Particle physicists at Ph.D.-granting institutions.* The overall results are strong and statistically significant. Again the null hypothesis that parameters of the model are jointly zero can be rejected. A-1 shows that productivity is significantly higher, by about 2.5 articles, in 1977 (T_3). The parameter estimates for the vintage variables imply that, compared to the group educated when field theory was in its prime (represented by the excluded vintage dummy), later vintages are less productive. This is consistent with evidence presented in the case study that those educated when field theory was important may have enjoyed an edge in particle physics. The differences, however, are only statistically significant for output measures PUB2 and PUB4, and then only between the field-theory group and the latest vintage (V_1), those receiving doctorates since 1970.

When individual fixed effects are not controlled for (model A), there is evidence of aging effects, a decline of 0.3 articles per period, much smaller, however, than was observed in solid-state physics. Model B shows that the null hypothesis that the unmeasurable individual fixed effects are jointly zero must be rejected. Moreover, once these effects are controlled for, there is no evidence to support the hypothesis that research activity declines over the life cycle. This outcome is not totally unexpected. The conceptual model implies that, when satisfaction from research is an argument in the utility function, research activity remains positive until retirement. In addition, the model suggests that the productivity profile is likely to be flatter when greater satisfaction is derived from puzzle-solving activity. Among the six groups studied, theoretical particle physicists most clearly fit this picture.

(ii) *Particle physicists at FFRDC's.* The results are fairly strong, although the null

hypothesis that the parameters of the model are jointly zero cannot be rejected in model A-2. As A-1 indicates, only in 1975 is there statistical support for the presence of time-period effects. Both estimations of model A, however, suggest that FFRDC particle physicists educated after 1957 are less productive than those educated before (about 4–10 articles less) and that the group educated after 1963 (V_1 and V_2) may be less productive than those educated between 1957 and 1963 (V_3). This is consistent with the case study's finding that the field-theory generation and those educated while field theory was still in vogue may have enjoyed an edge in particle physics.

As in the case of solid-state physics, there is evidence of a strong aging effect, a decline of 0.5 articles per period. Even after controlling for all individual-specific effects, research activity declines significantly with age. Apparently, particle physicists located at FFRDC's focus more heavily on the investment component of research than their peers at Ph.D.-granting institutions. One explanation for this is that more phenomenologists and experimentalists and fewer pure theorists are located at FFRDC's.

Atomic and Molecular Physics. Again the results are strong, although there is scant evidence of time-period effects. Output is significantly higher than the base period only for 1979 (T_4) for PUB1 and PUB3. Although the more-recent vintage is more productive than the earlier vintage, the difference is not statistically significant. This is consistent with the case study's observation that no major revolution has occurred since these scientists received their doctorates.

A statistically significant inverse relationship between age and publishing productivity is observed in A-1, (for PUB3, as well), although the confidence level is less than 0.95. This statistically significant age effect, however, does not hold up when the smaller data set of sample two is used to estimate model A. Just as in the previous fields, the null hypothesis of the absence of unmeasurable individual fixed effects is rejected. Controlling for these effects, however, as model B indicates, confirms the previous finding that productivity declines with age. (This is

particle physics. Our reading of the literature, however, and our correspondence and discussions with physicists do not lead us to believe that the choice of sector is endogenous to the model of publishing productivity in this case.

also true for PUB3). It appears (although only at a 0.90 confidence level) that output at first increases and then diminishes as the scientist ages, reaching a peak at age 39. (The peak for PUB3 is age 40).

B. Earth Science

The three areas studied are oceanography, geophysics, and geology. Oceanography relies heavily on geophysical theory and methods to investigate the oceans and lands beneath them; geophysics is the study of the earth, using the basic principles of physics; and geology focuses on how the earth was formed, its composition, history, and changes. In this study, geology includes the specialties of mineralogy, petrology, stratigraphy, sedimentation, paleontology, structural geology (tectonics), and geomorphology.

If there is one development over the past 50 years in earth science that has had the stature of a major conceptual change, it is clearly the revolutionary theory of a dynamic earth, called plate tectonics, developed in the mid-to-late 1960's. Doubt exists, however, as to whether the new plate-tectonic generation of scientists possesses a knowledge edge compared to their predecessors, particularly in geology and perhaps to a lesser extent in oceanography, where research activity often focuses on observation and classification, two activities which are thought to be unaffected by a major conceptual change.

Oceanography. The results are generally weak, and to some extent, this is undoubtedly attributable to the small sample size. The null hypothesis that the parameters of the model are jointly zero can only be rejected in model B. Time-period effects are not evident in model A, although they are present when output is adjusted for quality in PUB3 and PUB4. Compared to the earliest vintage (those receiving degrees prior to 1965), each later vintage in A-1 is, on average, less productive. (The null hypothesis of no difference must be accepted, however, when output is adjusted for coauthorship, PUB2.) Thus, it does not appear that those educated subsequent to the

plate-tectonic revolution gained a knowledge edge. The results also are not consistent with an alternative hypothesis, offered by one earth scientist, that those most likely caught up in the revolution would tend to be the most productive. The productivity differences between vintages, however, disappear when the vintage categories are compressed in A-2.

In A-1, publishing productivity is found to decline significantly with age; however, it declines by less than 0.1 article per period. (When the output measure is PUB3, the age variable is not statistically significant.) After controlling for all individual-specific fixed effects in model B, statistically significant life-cycle effects are also present. Output at first rises with age and then declines, and this decline begins very early in the career. (This is also true for PUB2, but productivity appears to decline linearly with age for PUB3 and PUB4).

Geophysics. The overall results for this field are strong. In all cases, the null hypothesis that the parameters of this model are jointly zero can be rejected at confidence levels exceeding 0.99. As model A shows, there are statistically significant positive differences (of 3-4 articles) in mean publishing rates in each time period compared to the base period, 1973. With one exception, all vintage effects are negative when compared to the earliest vintage, but only for the most recent vintage in the model for PUB4 is the difference statistically significant. These results do not support the latest-are-best-educated model of knowledge obsolescence. If anything, they suggest that the plate-tectonic generation has failed to keep pace with its predecessors.

Once again, there is evidence of an inverse relationship between publishing productivity and age (a decline of about 0.5 articles per period in A-1), and the results do not change appreciably in A-2. Furthermore, these life-cycle effects persist in model B. Output peaks quite late in the career, at age 59, and then declines. (The peak is 55 for PUB2, 58 for PUB3, and 53 for PUB4).

Geology. The empirical results are again generally strong in terms of the overall significance of the models. There does not,

however, appear to be a systematic pattern of differences over time in the mean level of publishing productivity, and, although the later vintages appear to be more productive than the earliest vintage, the differences are statistically significant only in A-2. This is not surprising. Geology is largely an observational field where vintage may be of little importance. As one author suggests (John Law, 1980 p. 160), "there is something about subjects such as geology which permits a conceptual pluralism that is relatively rare in physics or chemistry." Thus, even in the face of major revolutions in thought and practice, research may proceed in the usual manner.

Model A does not indicate the presence of aging effects. After controlling for all individual fixed effects in model B, however, aging effects do appear (a decline of 0.4 articles per period), although the coefficient is only significant at the 10-percent level.

IV. Conclusions

The major finding of this study is that, with the exception of particle physicists employed in Ph.D.-granting departments, life-cycle effects are present in a fully specified model of publishing productivity which, among other things, controls for individual fixed effects such as motivation and ability. Stated differently, there is evidence that, on average, scientists become less productive as they age. The aging effect that is found is attributed to age per se and not to the possibility that, for some reason, older scientists in the sample have different attributes, values, or access to resources than younger members of the sample. Hence, research activity over the life cycle appears to be investment-motivated.

The results, although tentative, also suggest that, for the most part, vintage matters, but not in the way predicted from a latest-educated-are-best-educated point of view. With the possible exception of geology, more recent vintages are never found to be significantly more productive than earlier vintages. Perhaps, in retrospect, this outcome is not all that surprising, given that the case studies suggest that, in at least some of the

fields, more-recent vintages may not have had a knowledge edge. For example, the physics case study suggests that atomic and molecular physics has not experienced dramatic changes in thought or technique during the past 40 years since the upheaval brought about by the quantum revolution. In other fields, such as solid-state/condensed-matter physics, although numerous developments could have produced vintage effects, it appears that there is a role for what one physicist called "ditch diggers," scientists who remain active by producing "backwater" research. Furthermore, the case study also suggests that, in particle physics, some later vintages may have enjoyed less of a knowledge edge because they were trained in concepts that subsequently proved to be dead ends. In addition, the earth-science case study raises doubt as to whether the plate-tectonic revolution as well as advances in computer technology would render older scientists coming from earlier vintages less productive.

There is, however, another more speculative explanation as to why the latest vintages, with the possible exception of geology, proved to be no more productive than the earlier vintages. During the 1960's and very early 1970's science grew very rapidly. It is possible that scientists obtaining doctorates during this period of rapid expansion may have been, on average, not as talented or motivated as scientists coming from earlier cohorts, which represent a smaller, more elite portion of the population. As a result, even if these scientists have a knowledge edge, a "talent deficit" may make them no more productive than their peers.

In the econometric model specified, a sample-selection variable was introduced to control for market conditions affecting employment location, and vintage dummies were introduced to capture knowledge-obsolescence effects. Since no variable directly controlled for the average ability or motivation of the cohort, it is possible that the vintage dummies reflect not only change in knowledge but also change in the average ability or motivation of the cohort. This would explain why the most-recent vintages

are never found to be more productive, with the possible exception of geology. The results also suggest that this ability/motivation argument might be extended well into the 1970's and 1980's, long after the growth in science peaked, perhaps because the best students were attracted into careers in law, business, and medicine. Bowen and Schuster (1986 pp. 224-6) document a marked shift away from the academic sector in the career choices of such highly talented members of the population as Rhodes scholars and Phi Beta Kappa members over the period 1965-1979.

Together, the aging and vintage results suggest that, during the next 10 or 15 years, the American scientific community will not be as productive as it was in the 1960's and early 1970's, assuming market conditions do not change dramatically. Not only will the community be older, but over time the community will become increasingly dominated by scientists who did not come from particularly productive cohorts.

APPENDIX A—DATA

Previous research on the relationship between age and scientific performance has been hampered by the lack of a comprehensive data base containing measures of productivity for scientists. We have assembled such a data base by using a computer algorithm to link the journal-publication data contained in the *Science Citation Index* (SCI) with the biennial *Survey of Doctorate Recipients* (SDR).

The number of journal articles is chosen as the measure of productivity, since it is generally recognized that the journal literature is the major outlet for recording scientific advances in many disciplines (Henry Menard, 1971). Publication counts are established for scientists trained, as designated in either the SDR or the *Survey of Earned Doctorates* as biochemists, earth scientists, physicists (excluding astronomers and astrophysicists), or physiologists by a computer algorithm using the source and corporate address files of the SCI. We count the flow of publications for a period of two years, beginning one year after the survey date of the scientist (1973, 1975, 1977, or 1979), given evidence on the length of the lag between the inception of the research project and the time at which the resulting output is likely to appear (Nelson and Pollock, 1970).

The magnitude of this project can be seen by considering the size of the two files that were linked. On the SDR side, even though we restricted the analysis to scientists in just four fields, including all sectors of employment there were 18,909 records for the 1973-1979 interview period. On the SCI side, over the

period 1974-1981, there were in excess of 9.6 million entries in the *Source Index*. Because of the confidential nature of the SDR, all work linking the data bases was performed by the Data Processing Unit at the *National Research Council* (NRC). The initial match procedure used was a variant of that developed by George Boyce at NRC for the study conducted by Lyle Jones et al. (1982). The match queued on last name, first name, middle initial (if present), state/country, and zip code. The completed match is approximately 95 percent accurate (Stephan and Levin, 1988).

A large part of this research involved the development of case studies to specify the market-determined dummy variables for cohort effects used in the selection equation, as well as the vintage dummy variables included in model A. These case studies are available upon request. Information for the case studies was gathered from various publications, including those produced by outside observers of the field such as historians and sociologists of science, personal interviews, and a mail survey. In interviews and in the questionnaire, scientists were asked to identify changes occurring in their specialty, either in theory or in research techniques, that could have negative effects on the productivity of persons trained before the innovation occurred. Two physicists, Steve Manson and Steve Sigur from Atlanta, gave particularly freely of their time, as did Spencer Weart of the American Institute of Physics (AIP). In addition, Beverly Porter of AIP was extremely helpful in identifying cohort effects. In physics, the survey was sent to 63 scientists, many of whom were members of the Brinkman Panels, established by NRC to review the state of physics in the 1980's. Of the 63, 26 physicists replied, either by completing the survey or by writing a letter. Of the 38 surveys mailed to geoscientists, eight returns were particularly useful. The late Bill Menard, a distinguished oceanographer, and William Glen, editor of *Eos*, were especially helpful.

APPENDIX B—ESTIMATION

The Olsen technique was used to obtain the sample-selection correction variable, SVAR. For each survey year, the probability of sample inclusion in a Ph.D.-granting department was estimated by ordinary least squares using the following regressors: the quality of graduate training, age, age-squared, whether the respondent was born in the South, whether the respondent was born in the non-South or Canada, the age at time of Ph.D., market-determined cohort effects (dummy variables), and interactions between the quality of graduate training and the market cohort effects. All of these variables, with the exception of categorical variables representing the quality of graduate training and dummy variables for the cohort effects are taken from the National Research Council's *Survey of Doctorate Recipients* (SDR). Data on the rankings of graduate departments over time (Hayward Keniston, 1959; Allan Cartter 1966; Kenneth Roose and Charles Anderson, 1970; Jones et al., 1982) are used to sort the Ph.D.-granting institutions into five categories, ranging from departments that were not ranked to departments

TABLE B1—DESCRIPTIVE STATISTICS FOR SUBFIELDS OF PHYSICS

Variable	Mean	SD	N
Solid-state physicists:			
PUB1 = total count of publications for two-year period	3.830	4.086	182
AGE = age	41.533	7.093	182
T_1 = 1 if year of survey = 1973; 0 otherwise	0.280	0.450	182
T_2 = 1 if year of survey = 1975; 0 otherwise	0.335	0.473	182
T_3 = 1 if year of survey = 1977; 0 otherwise	0.269	0.445	182
T_4 = 1 if year of survey = 1979; 0 otherwise	0.115	0.320	182
V_1 = 1 if year of Ph.D. > 1972; 0 otherwise	0.033	0.179	182
V_2 = 1 if 1963 ≤ year of Ph.D. ≤ 1972; 0 otherwise	0.473	0.501	182
V_3 = 1 if 1956 ≤ year of Ph.D. ≤ 1962; 0 otherwise	0.357	0.480	182
V_4 = 1 if 1948 ≤ year of Ph.D. ≤ 1955; 0 otherwise	0.099	0.299	182
V_5 = 1 if year of Ph.D. < 1948; 0 otherwise	0.039	0.193	182
VIN1 = 1 if year of Ph.D. > 1963; 0 otherwise	0.516	0.501	159
VIN2 = 1 if 1956 ≤ year of Ph.D. ≤ 1962; 0 otherwise	0.365	0.483	159
VIN3 = 1 if year of Ph.D. < 1956; 0 otherwise	0.119	0.325	159
Particle physicists at Ph.D.-granting institutions:			
PUB1 = total count of publications for two-year period	2.982	2.903	168
AGE = age	39.786	6.841	168
T_1 = 1 if year of survey = 1973; 0 otherwise	0.280	0.450	168
T_2 = 1 if year of survey = 1975; 0 otherwise	0.333	0.473	168
T_3 = 1 if year of survey = 1977; 0 otherwise	0.286	0.453	168
T_4 = 1 if year of survey = 1979; 0 otherwise	0.101	0.303	168
V_1 = 1 if year of Ph.D. > 1970; 0 otherwise	0.143	0.351	168
V_2 = 1 if 1964 ≤ year of Ph.D. ≤ 1970; 0 otherwise	0.387	0.489	168
V_3 = 1 if 1957 ≤ year of Ph.D. ≤ 1963; 0 otherwise	0.333	0.473	168
V_4 = 1 if year of Ph.D. < 1957; 0 otherwise	0.137	0.345	168
Particle physicists at FFRDC's:			
PUB1 = total count of publications for two-year period	2.682	3.843	157
AGE = age	39.433	6.841	157
T_1 = 1 if year of survey = 1973; 0 otherwise	0.248	0.434	157
T_2 = 1 if year of survey = 1975; 0 otherwise	0.299	0.459	157
T_3 = 1 if year of survey = 1977; 0 otherwise	0.312	0.465	157
T_4 = 1 if year of survey = 1979; 0 otherwise	0.140	0.348	157
V_1 = 1 if year of Ph.D. > 1970; 0 otherwise	0.172	0.379	157
V_2 = 1 if 1964 ≤ year of Ph.D. ≤ 1970; 0 otherwise	0.440	0.498	157
V_3 = 1 if 1957 ≤ year of Ph.D. ≤ 1963; 0 otherwise	0.299	0.459	157
V_4 = 1 if year of Ph.D. < 1957; 0 otherwise	0.089	0.286	157
Atomic and molecular physicists:			
PUB1 = total count of publications for two-year period	3.258	4.368	89
AGE = age	44.371	9.048	89
T_1 = 1 if year of survey = 1973; 0 otherwise	0.258	0.440	89
T_2 = 1 if year of survey = 1975; 0 otherwise	0.303	0.462	89
T_3 = 1 if year of survey = 1977; 0 otherwise	0.281	0.452	89
T_4 = 1 if year of survey = 1979; 0 otherwise	0.157	0.366	89
V_1 = 1 if year of Ph.D. > 1963; 0 otherwise	0.416	0.496	89
V_2 = 1 if year of Ph.D. ≤ 1963; 0 otherwise	0.584	0.496	89

that were in the top five. These intermediate results are available upon request.

Our commitment to a fixed-effects specification necessitated restricting the analysis to scientists employed in the top sector more than once. This meant excluding approximately 17 percent of the observations. The predominant reason that scientists appeared in the top sector only once was because they were surveyed (or responded) only once. To investigate the possibility

that other scientists systematically self-selected out of the top sector, following Daniel Hamermesh (1987), we tested whether the structure of model A differed between those who were in the top sector only once and those who were in more than once. On the basis of likelihood-ratio tests, the null hypothesis of no difference in the basic structure of model A for the two groups could not be rejected at the 95-percent level for all measures of output in all subfields.

TABLE B2—DESCRIPTIVE STATISTICS FOR SUBFIELDS OF EARTH SCIENCE

Variable	Mean	SD	N
Oceanographers:			
PUB1 = total count of publications for two-year period	2.105	1.839	57
AGE = age	40.333	7.666	57
T_1 = 1 if year of survey = 1973; 0 otherwise	0.211	0.411	57
T_2 = 1 if year of survey = 1975; 0 otherwise	0.281	0.453	57
T_3 = 1 if year of survey = 1977; 0 otherwise	0.281	0.453	57
T_4 = 1 if year of survey = 1979; 0 otherwise	0.228	0.423	57
V_1 = 1 if year of Ph.D. > 1969; 0 otherwise	0.298	0.462	57
V_2 = 1 if 1965 ≤ year of Ph.D. ≤ 1969; 0 otherwise	0.579	0.498	57
V_3 = 1 if year of Ph.D. < 1965; 0 otherwise	0.123	0.331	57
VIN1 = 1 if year of Ph.D. > 1969; 0 otherwise	0.275	0.451	51
VIN2 = 1 if year of Ph.D. ≤ 1969; 0 otherwise	0.725	0.451	51
Geophysicists:			
PUB1 = total count of publications for two-year period	3.654	4.404	78
AGE = age	41.474	8.208	78
T_1 = 1 if year of survey = 1973; 0 otherwise	0.218	0.416	78
T_2 = 1 if year of survey = 1975; 0 otherwise	0.282	0.453	78
T_3 = 1 if year of survey = 1977; 0 otherwise	0.282	0.453	78
T_4 = 1 if year of survey = 1979; 0 otherwise	0.218	0.416	78
V_1 = 1 if year of Ph.D. > 1969; 0 otherwise	0.321	0.470	78
V_2 = 1 if 1965 ≤ year of Ph.D. ≤ 1969; 0 otherwise	0.180	0.386	78
V_3 = 1 if 1960 ≤ year of Ph.D. ≤ 1964; 0 otherwise	0.282	0.453	78
V_4 = 1 if 1955 ≤ year of Ph.D. ≤ 1959; 0 otherwise	0.115	0.322	78
V_5 = 1 if year of Ph.D. < 1955; 0 otherwise	0.103	0.305	78
VIN1 = 1 if year of Ph.D. > 1969; 0 otherwise	0.333	0.475	69
VIN2 = 1 if 1965 ≤ year of Ph.D. ≤ 1969; 0 otherwise	0.203	0.405	69
VIN3 = 1 if 1960 ≤ year of Ph.D. ≤ 1964; 0 otherwise	0.290	0.457	69
VIN4 = 1 if year of Ph.D. < 1960; 0 otherwise	0.058	0.235	69
Geologists:			
PUB1 = total count of publications for two-year period	1.535	2.087	172
AGE = age	47.500	10.245	172
T_1 = 1 if year of survey = 1973; 0 otherwise	0.273	0.447	172
T_2 = 1 if year of survey = 1975; 0 otherwise	0.291	0.455	172
T_3 = 1 if year of survey = 1977; 0 otherwise	0.244	0.431	172
T_4 = 1 if year of survey = 1979; 0 otherwise	0.192	0.395	172
V_1 = 1 if year of Ph.D. > 1970; 0 otherwise	0.134	0.341	172
V_2 = 1 if 1965 ≤ year of Ph.D. ≤ 1969; 0 otherwise	0.081	0.274	172
V_3 = 1 if 1960 ≤ year of Ph.D. ≤ 1964; 0 otherwise	0.244	0.431	172
V_4 = 1 if 1955 ≤ year of Ph.D. ≤ 1959; 0 otherwise	0.204	0.404	172
V_5 = 1 if 1945 ≤ year of Ph.D. ≤ 1954; 0 otherwise	0.192	0.395	172
V_6 = 1 if year of Ph.D. < 1945; 0 otherwise	0.145	0.354	172

In addition to age, vintage, the correction for sample-selectivity bias (SVAR), and time-period dummy variables, model A also controls for REPRANK, SALARY, ADMIN/TEACH, and FEDSUP. REPRANK is taken from Jones et al. (1982), while the other variables are derived from data in the SDR. REPRANK measures the reputational rating of graduate departments and is included given the abundance of evidence that productivity is positively related to department quality (Long, 1978; Long and McGinnis, 1981; Fox, 1983). Moreover, it is also highly correlated with other measures of the richness of the research environment in which the scientist works. SALARY, the scientist's adjusted (for inflation) annual salary, is related to the scientist's past productivity and hence

serves as a proxy for reputational prestige and cumulative advantage. ADMIN/TEACH captures whether the scientist devotes considerable effort to nonresearch activities. Finally, FEDSUP indicates whether the scientist presently has government research support. Although not reported in the text, the findings with respect to all control variables were generally consistent with expectations.

The preliminary analysis also considered additional explanatory variables. These included such proxies for ability as whether the individual attended a select undergraduate institution, the rating of the graduate institution attended, and proxies (dummy variables) to capture exogenous differences in publishing productivity by field because scientists may be working in fields

other than their field of training. Since in no case did it appear that these additional regressors had a marked effect on the results, they have been omitted from the findings reported here. Also considered in preliminary work was a dummy variable for tenure status and a variable for the number of years since tenure was received. Unfortunately, because there were so many cases in which the scientist failed to indicate tenure status and numerous inconsistencies between the years since tenure and the reported employment history, analysis including these variables was not pursued further.

REFERENCES

- Allison, Paul D. and Stewart, John A., "Productivity Differences Among Scientists: Evidence for Accumulative Advantage," *American Sociological Review*, August 1974, 39, 596-606.
- _____, Long, J. Scott and Krauze, Tad K., "Cumulative Advantage and Inequality in Science," *American Sociological Review*, October 1982, 47, 615-25.
- Alpher, R. A., Fiske, M. D., Ham, F. S. and Kahn, P. B., "Summary of a Statistical Study of the Ph.D. Physicist Employed in Industry," in *The Transition in Physics Doctoral Employment, 1960-1990*, Report of the Physics Manpower Panel of the American Physical Society, M. D. Fiske, Chairman, New York: American Institute of Physics, August 1979, 25-32.
- Bayer, Alan E. and Dutton, Jeffrey E., "Career Age and Research-Professional Activities of Academic Scientists," *Journal of Higher Education*, May/June 1977, 48, 259-82.
- Ben-Porath, Yoram, "The Production of Human Capital and the Life Cycle of Earnings," *Journal of Political Economy*, August 1967, 75, 352-65.
- Bils, Mark, "Real Wages Over the Business Cycle: Evidence from Panel Data," *Journal of Political Economy*, August 1985, 93, 666-89.
- Bowen, Howard R. and Schuster, Jack, *American Professors: A National Resource Imperiled*, New York: Oxford University Press, 1986.
- Cartter, Allan M., *An Assessment of Quality in Graduate Education*, Washington, DC: American Council on Education, 1966.
- Cole, Gerald A., "Classifying Research Units by Patterns of Performance and Influence: A Typology of Round I Data," in Frank Andrews, ed., *Scientific Productivity*, Cambridge: Cambridge University Press, 1979, 353-404.
- Cole, Stephen, "Age and Scientific Performance," *American Journal of Sociology*, January 1979, 84, 958-77.
- Diamond, Arthur M., Jr., "An Economic Model of the Life-Cycle Research Productivity of Scientists," *Scientometrics*, 1984, 6 (3), 189-96.
- _____, "The Life-Cycle Research Productivity of Mathematicians and Scientists," *Journal of Gerontology*, July 1986, 41, 520-5.
- Fox, Mary F., "Publication Productivity Among Scientists: A Critical Review," *Social Studies of Science*, May 1983, 13, 283-305.
- Garfield, Eugene, ed., *SCI Journal Citation Reports*, Philadelphia, PA: Institute for Scientific Information, 1976.
- Hagstrom, Warren O., *The Scientific Community*, New York: Basic Books, 1965.
- Hamermesh, Daniel S., "Why Do Fixed-Effects Models Perform So Poorly? The Case of Academic Salaries," National Bureau of Economic Research, Working Paper No. 2135, January 1987.
- Hausman, Jerry A., "Specification Tests in Econometrics," *Econometrica*, November 1978, 46, 1251-71.
- _____, and Taylor, William E., "Panel Data and Unobserved Individual Effects," *Econometrica*, November 1981, 49, 1377-98.
- Heckman, James J., "Sample Selection Bias as a Specification Error," *Econometrica*, January 1979, 47, 153-61.
- _____, and Macurdy, Thomas E., "A Life Cycle Model of Female Labor Supply," *Review of Economic Studies*, January 1980, 47, 47-74.
- Johnson, William R., "Vintage Effects in the Earnings of White American Men," *Review of Economics and Statistics*, August 1980, 62, 399-407.
- Jones, Lyle V., Lindsey, Gardner and Coggeshall, Porter, eds., *An Assessment of Research Doctorate Programs in the U.S.: Mathematical and Physical Sciences*, Washing-

- ton, DC: National Academy Press, 1982.
- Judge, George G., Griffiths, William E., Hill, R. Carter and Lee, Tsoung-Chao, *The Theory and Practice of Econometrics*, New York: Wiley, 1980.
- Keniston, Hayward, *Graduate Study and Research in the Arts and Sciences at the University of Pennsylvania*, Philadelphia: University of Pennsylvania Press, 1959.
- Law, John, "Fragmentation and Investment in Sedimentology," *Social Studies of Science*, February 1980, 10, 1-22.
- Lehman, Harvey C., *Age and Achievement*, Princeton, NJ: Princeton University Press, 1953.
- Long, J. Scott, "Productivity and Academic Position in the Scientific Career," *American Sociological Review*, December 1978, 43, 889-908.
- _____ and McGinnis, Robert, "Organizational Context and Scientific Productivity," *American Sociological Review*, August 1981, 46, 422-42.
- Maddala, G. S., *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press, 1983.
- _____, "Limited Dependent Variable Models Using Panel Data," *Journal of Human Resources*, Summer 1987, 22, 307-38.
- McDowell, John M., "Obsolescence of Knowledge and Career Publication Profiles: Some Evidence of Differences Among Fields in Costs of Interrupted Careers," *American Economic Review*, September 1982, 72, 752-68.
- Menard, Henry W., *Science: Growth and Change*, Cambridge, MA: Harvard University Press, 1971.
- Miller, A. Carolyn and Serzan, Sharon L., "Criteria for Identifying a Refereed Journal," *Journal of Higher Education*, November/December 1984, 6, 673-97.
- Mincer, Jacob, *Schooling, Experience, and Earnings*, New York: Columbia University Press (for the National Bureau of Economic Research), 1974.
- Nelson, Carnot E. and Pollock, Donald K., eds., *Communications Among Scientists and Engineers*, Lexington, MA: D. C. Heath—Lexington Books, 1970.
- Olsen, Randall J., "A Least Squares Correction for Selection Bias," *Econometrica*, November 1980, 48, 1815-20.
- Polachek, Solomon, "Occupational Segregation: An Alternative Hypothesis," *Journal of Contemporary Business*, Winter 1976, 5, 1-12.
- Pollak, Robert A. and Wales, Terrence J., "Estimation of the Linear Expenditure System," *Econometrica*, October 1969, 37, 611-28.
- Porter, Beverly Fearn, (1979a) "Mobile Young Faculty: A Follow-Up Study of Untenured Assistant Professors Leaving a Sample of Top Physics Departments," in *The Transition in Physical Doctoral Employment, 1960-1990*, Report of the Physics Manpower Panel of the American Physical Society, M. D. Fiske, Chairman, New York: American Institute of Physics, August 1979, 49-112.
- _____, (1979b) "Transition: A Follow-Up Study of 1973 Postdoctorals," in *The Transition in Physics Doctoral Employment, 1960-1990*, Report of the Physics Manpower Panel of the American Physical Society, M. D. Fiske, Chairman, New York: American Institute of Physics, August 1979, 113-92.
- Rodgers, Willard L., "Estimable Functions of Age, Period, and Cohort Effects," *American Sociological Review*, December 1982, 47, 774-87.
- Roose, Kenneth D. and Andersen, Charles J., *A Rating of Graduate Programs*, Washington, DC: American Council on Education, 1970.
- Ryder, Harl E., Stafford, Frank P. and Stephan, Paula E., "Labor, Leisure and Training over the Life Cycle," *International Economic Review*, October 1976, 17, 651-74.
- Stephan, Paula E., "Human Capital Production: Life-Cycle Production with Different Learning Technologies," *Economic Inquiry*, December 1976, 14, 539-57.
- _____ and Levin, Sharon G., "Sex Segregation in Education: The Case of Doctorate Recipients," *Journal of Behavioral Economics*, Winter 1983, 12, 67-94.
- _____ and _____, "Age-Publishing Patterns in Science: Estimation and Measurement Issues," in A. F. J. van Raan, ed., *A Handbook of Quantitative Studies*

- of Science and Technology*, Amsterdam: Elsevier-North Holland, 1988, 31-80.
- Tobin, James**, "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, January 1958, 26, 24-36.
- Zuckerman, Harriet**, *Scientific Elite*, New York: The Free Press, 1977.
- National Research Council**, *Research Excellence Through the Year 2000: The Importance of Maintaining a Flow of New Faculty into Academic Research*, Washington, DC: National Academy of Sciences, 1979.
- _____, *Survey of Doctorate Recipients*, biennial.
- _____, *Survey of Earned Doctorates*, annual.
- National Science Foundation**, *Doctoral Scientists and Engineers: A Decade of Change*, Washington, DC: National Science Foundation, 1988, 88-302.
- Science Citation Index**, Philadelphia, PA: Institute for Scientific Information, annual.

Economic Effects of Federal Credit Programs

By WILLIAM G. GALE*

Since 1980, the federal government has directly subsidized one-third of all nonfederal borrowing. This paper presents numerical estimates of the effects of federal lending. Existing credit subsidies appear to have important effects on the allocation of credit, but little effect on aggregate investment. Efficiency costs are shown to be large (approximately 1/3 percent of GNP). Government costs exceed 50 cents per dollar of incremental targeted lending. Interactions among programs can eliminate much or all of the original gain provided by a subsidy, especially if borrowers are rationed. The paper also examines the effects of several policy reforms. (JEL 313, 321)

From 1980 to 1987, the Federal government borrowed \$1,253 billion to cover budgetary shortfalls.¹ Large and persistent fiscal deficits have generated widespread attention in academic and policy debates. Over the same period, federal lending programs extended \$1,208 billion of net credit, while new federal and federally assisted lending exceeded \$4 trillion. Since 1980, the federal government has subsidized, guaranteed, or directly extended more than one-third of all borrowing by nonfederal sectors. However, despite the potential importance of federal credit activity, there exist few systematic analyses of the effects and appropriate role of the government as a lender. Barry Bosworth et al. (1987) provide a comprehensive discussion of issues and characteristics of lending programs but provide little formal analysis. Gregory Mankiw

(1986), David de Meza and David Webb (1987, 1988), Gale (1989, 1990b), and Bruce Smith and Michael Stutzer, (1989) provide theoretical models of the effects of credit subsidies in markets with imperfect information.² Charles Calomiris et al. (1986) analyze the empirical effects of agricultural credit policy.

The purpose of this paper is to develop a framework in which to analyze the quantitative effects of federal lending on credit allocation, economic efficiency, and related issues. The underlying model is based on Joseph Stiglitz and Andrew Weiss (1981) and posits asymmetric information between borrowers and lenders. Government interventions are modeled as part of an overall market for credit, which may be characterized by market clearing, rationing, or redlining of different borrower groups. Thus, it is possible to see how the effects of credit interventions depend upon the initial regime. In addition, the model accounts for the financing costs of federal credit, interactions among programs, and the use of alternative credit instruments.

The main contribution of this paper is to simulate the model described above to generate numerical estimates of the effects of

*Department of Economics, UCLA, Los Angeles, CA 90024-1477. This paper is based on my Ph.D. dissertation. I thank my advisor, Michael Boskin, for many helpful discussions. I have also benefitted from comments from Doug Bernheim, David Butz, Al Harberger, Frank Howland, Seonghwan Oh, Eric Rasmusen, John Riley, Russell Roberts, John Karl Scholz, John Shoven, Jonathan Skinner, Michael Waldman, and two anonymous referees, and I gratefully acknowledge financial support from the John M. Olin Foundation. I thank Lorraine Grams for expert assistance in the preparation of this paper.

¹All statistics in this paragraph are based on the data provided in Table 1, except for new federal and federally assisted lending, which is based on data provided in "Special Analysis F" of the U.S. Budget.

²Bruce Smith (1983) discusses optimal government lending with imperfect information but focuses on monetary policy rather than targeted credit subsidies. An important earlier paper on credit subsidies is Rudolph Penner and William Silber (1973).

credit programs. Principal conclusions are as follows. Existing credit subsidies raise aggregate private investment by between 0 percent and 4 percent, depending on the elasticity of the supply of funds. Crowding out depends critically on the supply elasticity but is bounded at a maximum of about 5 percent of original borrowing by nontargeted sectors. The allocational effects of lending policies depend on the size of the effective subsidy, rather than on credit volume. As a consequence, lending programs exert important effects on the allocation of funds to farmers, students, small businesses, and tax-exempt borrowers. The well-known federal mortgage-guarantee programs are estimated to have smaller net effects because of the smaller effective subsidies they provide. The allocational effects can also depend on the initial state of the market. If a target group is rationed, credit programs can exert very strong effects, by releasing the rationing constraint.

Although the underlying model allows for the possibility of welfare-improving interventions, the estimated efficiency cost of actual credit policy is high: approximately \$10–15 billion, or 1/3 percent of gross national product in 1987. These estimates are relatively insensitive to assumptions about the elasticity of supply of funds. Therefore, even if the crowding-out effects are small, the welfare loss is still sizable.

Because both new and inframarginal borrowers receive funds, credit subsidies cost the government in excess of 50 cents per dollar of *incremental* target-group investment. In addition, most direct welfare gains appear to accrue to borrowers who would have received credit without government subsidies. These subsidies represent pure windfall gains for the recipients, with no obvious corresponding societal benefits. All current programs require the existence of large external benefits to be welfare-improving. Even programs for groups that would have been excluded from private markets without federal assistance (say, students) face this requirement, because the default and subsidy costs are so large.

Because a subsidy to one target group crowds out other target groups as well as

nontargeted groups, interactions among programs can indirectly eliminate much or all of the original direct gain. In the low-supply-elasticity case examined here, up to 15 percent of the direct gains to farmers, students, and small businesses are offset by the effects of other credit programs. Interactions can be even more severe if a target group is rationed or redlined.

Two general principles emerge for policy reform. First, the effectiveness of any proposed reform depends on the structure of the original subsidy. For example, the vast majority of the cost of student-loan guarantees stems from the government's willingness to cover interest payments until the student graduates. Many policy proposals, however, focus on reducing the default rate, which would have only a minor impact on the costs of the program. Second, programs in which the government provides appropriate marginal incentives appear to be more efficient than programs that attempt to replace private credit arrangements. These results indicate that, although there is still a role for government in the more marginal sectors of the credit market, that role is more limited and somewhat different in nature than current policies would suggest.

Section I provides an overview of current federal credit activity. Section II develops the formal model. Section III specifies parameter values. Section IV presents the main results. A concluding section discusses caveats and extensions to this line of research. Appendixes A and B describe some important calculations and provide detailed references for the parameter specifications, respectively.

I. An Overview of Federal Credit Subsidies

The government supplies credit through several instruments.³ Direct loans typically offer large subsidies and are concentrated in the agricultural and rural sectors. Loan

³Detailed information concerning federal credit programs may be found in "Special Analysis F," in *Special Analyses: Budget of the U.S. Government* (any year) or in Bosworth et al. (1987).

TABLE 1—FEDERAL LENDING AND DOMESTIC CREDIT MARKETS (DOLLAR AMOUNTS IN BILLIONS)

Statistic	Fiscal Year								Total
	1980	1981	1982	1983	1984	1985	1986	1987	
Net credit advanced in nonfinancial credit markets	341.7	375.9	388.9	550.2	753.9	854.8	831.7	685.2	4,782.3
Net federal and federally assisted lending	110.2	109.9	131.8	140.2	129.9	246.7	159.9	179.5	1,208.1
Direct loans	24.2	26.1	23.4	15.3	6.3	28.0	11.2	-19.0	115.5
Guaranteed loans	31.6	28.0	20.9	34.1	20.1	21.6	34.6	60.4	251.3
GSE	24.1	32.4	43.3	37.1	53.1	60.7	83.3	107.8	441.8
Tax-exempt credit	30.3	23.4	44.2	53.7	50.4	136.4	30.8	30.3	399.5
Federal borrowing from the public	70.5	79.3	135.0	212.3	170.8	197.3	236.3	151.7	1,253.2
Federal lending as a percentage net credit	32	29	34	25	17	29	19	26	25
Federal lending as a percentage of federal borrowing	156	138	98	66	76	125	68	118	96
Federal lending as a percentage of nonfederal borrowing	41	37	52	41	22	38	27	34	34

Note: Net lending is new lending minus repayments.

Sources: Office of Management and Budget, "Special Analysis F," *Special Analyses: Budget of the United States*, 1988 table F-22; Board of Governors of the Federal Reserve System, *Flow of Funds Accounts*, Fourth Quarter, 1987, pp. 2-3.

guarantees assist a wide range of borrowers, particularly home-owners, small businesses, students, and exporters. Clifford Hardin and Arthur Denzau (1981) report the existence of more than 350 direct and guaranteed lending programs. Government-sponsored enterprises (GSE) aid borrowers in housing, agricultural, and student loan markets, primarily through the operation of secondary markets.⁴ Tax-exempt status allows state and local governments to borrow at reduced cost and to operate their own credit programs by passing on the interest savings to preferred borrowers. Effective subsidy rates, default rates, and credit volume vary greatly across sectors and programs and are discussed in detail below.

Credit policies are typically justified in terms of preexisting distortions (e.g., student loans), distributional concerns (farm loans), or externalities (small businesses). Credit subsidies have also been proposed as solutions to issues such as energy development and the maintenance of international

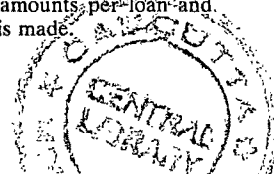
competitiveness. In addition, emergency loan guarantees have assisted Chrysler, Lockheed, and New York City in recent years.⁵

As shown in Table 1, lending programs account for an important share of credit advanced in domestic capital markets. The volume of net federal lending has grown at approximately the same rate as overall credit in the 1980's and has exceeded \$100 billion in every year since 1980. In a typical year, federal lending accounts for one-fourth of all net credit advanced and one-third of all nonfederal borrowing. Since 1980, the amount of net federal and federally assisted lending is approximately the same as the amount of federal borrowing.

Extensive lending activity, per se, does not imply that credit policy imparts important effects on the economy. However, the sheer magnitude and pervasiveness of credit interventions suggest the need for further

⁴For a detailed discussion of issues concerning government-sponsored enterprises see Thomas Stanton (1988).

⁵Paul Chaney and Anjan Thakor (1985) analyze emergency guarantees. However, these should be distinguished from ongoing guarantee programs that guarantee many loans and smaller amounts per loan and that do so at the time the loan is made.



analysis. In particular, the volume of federal credit activity indicates that any formal analysis should consider the effects of subsidies on nontargeted groups and the financing of credit policy. The diversity indicates the need to model several types of lending instruments and interactions among credit programs. Finally, since many of the targeted groups appear to be the marginal borrowers in modern credit markets, analysis of both rationing and market-clearing outcomes will be relevant. The next section outlines a model with these characteristics.

II. A Model of the Credit Market⁶

The underlying model is a variant of Stiglitz and Weiss (1981) and analyzes a competitive loan market in which lenders have imperfect information about borrowers. The market consists of government and many borrowers, depositors, and financial intermediaries. All agents are risk-neutral; there is no aggregate risk. Depositors supply funds to the market according to a function $S(\rho)$, where ρ is the certain rate of return on bank deposits.

Borrowers are divided into $n+1$ groups: n target groups for credit policy and one general (nontargeted) group.⁷ Borrowers are characterized by two pieces of information: their group identity and their location within that group. Group identity is assumed to be public information and will provide a (noisy) signal of borrowers' riskiness. Location within a group refers to the riskiness of the individual's projects and is known only to the individual borrower. With these informational assumptions, lenders (including the government) can determine which groups are eligible for credit subsidies, yet still face residual uncertainty within each group.

Borrower groups are indexed by $i = 0, 1, \dots, n$ and characterized by two behavioral assumptions. First, the demand for

loans (L_i^D) is a decreasing function of the effective interest rate (r_i^*) paid by the borrower. Second, the repayment rate (ϕ_i) falls as r_i^* rises. That is, increases in effective interest rates generate adverse selection in each loan market.⁸

Financial intermediation ("banking") is subject to free entry and constant returns to scale. Since they can discern only a noisy signal of borrowers' riskiness, banks set different loan rates for each group.⁹ The expected return to the bank on a loan to target group i at rate r_i is

$$\rho_i = \rho_i(r_i, \phi_i, C_i)$$

where C_i (a vector) represents credit policy for target group i . Holding credit policy constant, the existence of adverse selection implies that an increase in lending rates may eventually lead to a reduction in the banks' expected return.¹⁰ This relationship is shown as the ρ_j curve in Figure 1.

⁸These results can easily be derived from primitive assumptions on information, endowments, technology, and preferences. For example, Gale (1990b) uses the information structure above and assumes that an investor in group I with riskiness j has no endowment, must borrow to invest, has an available project of unit size which succeeds with probability $p_I(j)$, pays $R_I(j)$ if successful, and pays 0 otherwise. For any two borrowers j_1 and j_2 in group I , if $R_I(j_1) > R_I(j_2)$ then $p_I(j_1) < p_I(j_2)$. This implies that

$$L_I^D(r_I^*) = \int_0^{j_I^*(r_I^*)} f_I(j) dj$$

and

$$\phi_I(r_I^*) = \left[\int_0^{j_I^*(r_I^*)} p_I(j) f_I(j) dj \right] / F_I(j_I^*(r_I^*))$$

where j_I^* is the marginal borrower in group j (i.e., j_I^* is defined by $R_I(j_I^*) = r_I^*$) and where f and F represent the density and distribution of j , respectively. Similar specifications can be used to generate the supply curve. Thus, the underlying supply and demand functions are structural (i.e., invariant with respect to credit policy).

⁹Allowing banks to set an interest and collateral requirement for each group would not change the results as long as the collateral is incomplete (i.e., as long as banks prefer that borrowers repay rather than default).

¹⁰More generally, adverse selection implies that ρ need not be monotonic in r (see Stiglitz and Weiss, 1981).

⁶This section provides a brief overview of the underlying theoretical model, which is derived rigorously in Gale (1990b).

⁷Additional nontargeted groups can be incorporated as well.

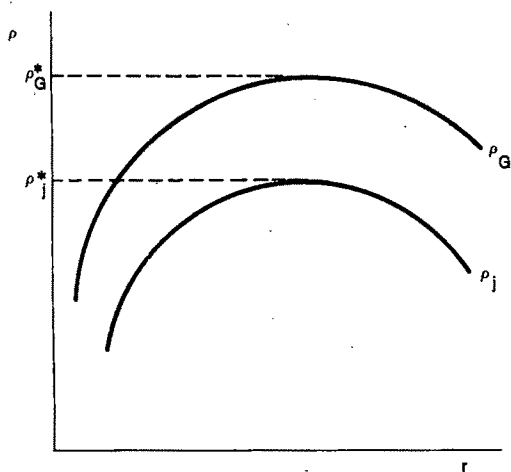


FIGURE 1. BANK RETURN ON GENERAL LOANS (ρ_G) AND TARGET-GROUP LOANS (ρ_j)

As shown in Figure 1, the maximum bank return on general loans (ρ_G^*) is assumed to be at least as great as that available on target group loans: $\rho_G^* \geq \rho_j^*$, $j = 1, \dots, n$.¹¹ This assumption will prove necessary and sufficient to imply that situations exist in which the general market clears while target groups are rationed or redlined.

The government may assist target groups through subsidized lending, loan guarantees, or tax-exempt status. The subsidies reduce the effective interest rate and raise the expected bank return to target-group lending indirectly through $\phi_i(r_i^*)$ and in some cases directly (e.g., loan guarantees). The government borrows to fund its credit programs and pays its creditors with program revenue first and lump-sum taxes on depositors. These assumptions force explicit recognition of the costs of the programs. The government has the same information and borrowing costs as banks.

The model is competitive in that there are many agents and entry is free. There is, however, one important departure from the standard competitive framework: in addition to caring about aggregate demand,

¹¹Figure 1 also shows $\rho_G(r) > \rho_j(r) \forall j, r$. This assumption is not necessary for any of the results, but simplifies the figure.

banks also care about the identity of the borrowers demanding loans. Thus, an important concept will be effective demand—demand for loans that banks are willing to supply. Effective demand for borrower group i is given by

$$L_i = \begin{cases} L_i^D & \text{if } \hat{\rho} < \rho_i^*(r_i^*) \\ 0 & \text{if } \hat{\rho} > \rho_i^*(r_i^*) \end{cases}$$

where $\hat{\rho}$ is the equilibrium value of ρ . For example, in Figure 1, if the equilibrium cost of funds is greater than ρ_j^* , banks will not consider making loans to group j , since such loans would generate negative expected profits. If $\hat{\rho} = \rho_i^*$, L_i is determined as a residual, after other credit demands have been met, such that $0 \leq L_i \leq L_i^D(r_i^*)$. These considerations imply that aggregate effective demand, shown in Figure 2 (for $n = 1$), is a step function.¹² In equilibrium, the market for target-group loans may be characterized by market clearing (S_3), rationing (S_2), or redlining (S_1), depending on the relative magnitudes of supply and effective demand.¹³

Equilibrium is characterized by two economic conditions. First, banks' expected returns to lending are equalized across all groups that receive loans and are equal to the cost of funds (zero profits):

$$(1) \quad \hat{\rho} = \rho_i \quad \text{if } \rho_i^* \geq \hat{\rho}.$$

Second, the sum of private and public demands equals the supply of funds:

$$(2) \quad S(\rho) = \sum_{i=0}^n [L_i + G_i]$$

where G_i is government borrowing to fund programs for group i . A third set of equations shows that credit subsidies drive a

¹²See John Riley (1987) for an independent derivation of this demand curve.

¹³A supply curve intersecting the horizontal demand strip to the left of S_1 would represent rationing of general borrowers and redlining of target borrowers. This case is not considered.

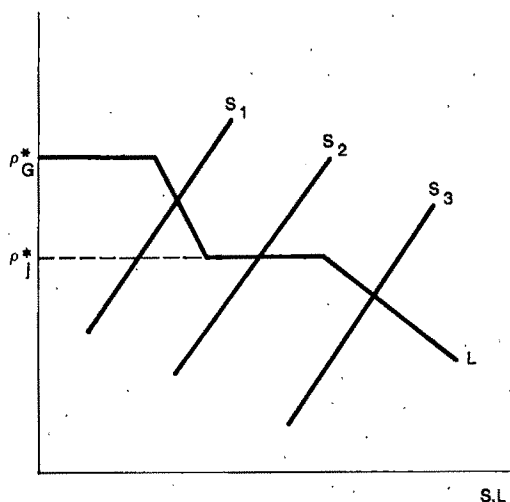


FIGURE 2. AGGREGATE EFFECTIVE DEMAND FOR $n=1$

wedge between effective interest rates and loan rates:

$$(3) \quad r_i^* = g(r_i, C_i) \quad i = 0, 1, \dots, n.$$

The simulations employ the following supply and demand functions:

$$(4) \quad S = a(\rho)^\alpha$$

$$(5) \quad L_i^D = b_i(r_i^*)^{\beta_i} \quad i = 0, 1, \dots, n$$

where α and β_i represent constant elasticities and a and b_i are constants. As shown in Appendix A, government borrowing to fund credit policy for group i is given by

$$(6) \quad G_i = L_i s_i$$

where s_i is the percentage reduction in the present discounted value of borrowers' loan payments due to federal credit subsidies.

Appendix A also shows that the relationship among ρ , r_i , and r_i^* may be reduced to a single set of equations, summarized here as¹⁴

$$(7) \quad r_i^* = f(\rho, \phi_i, C_i) \quad i = 0, 1, \dots, n.$$

¹⁴The loan rate, r_i , vanishes from the equilibrium calculation because it is determined by ρ , ϕ_i , and C_i .

The effective interest rate rises with increases in deposit rates and falls with increases in credit subsidies and the probability of repayment. Equations (2) and (4)–(7) represent the market-clearing equilibrium, with S , L_i^D , G , ρ , and r_i^* endogenous.¹⁵

If group j is rationed, equation (2) is modified to become

$$(8) \quad L_j + G_j = S - \sum_h (L_h^D + G_h)$$

where h indexes groups with $\rho_h^* > \hat{\rho}$. Lending to group j is determined as a residual. As shown in Figure 2, the equilibrium cost of funds must equal the maximum bank rate of return on group- j loans:

$$(9) \quad \hat{\rho} = \rho_j^*.$$

Equations (4)–(9) represent the rationing equilibrium.

One important characteristic of the model is that, in equilibrium, banks order borrower groups by their maximum rate of return and serve the groups sequentially. If the equilibrium cost of funds is $\hat{\rho}$, all groups j with $\rho_j^* > \hat{\rho}$ have clearing credit markets, those with $\rho_j^* = \hat{\rho}$ are rationed,¹⁶ and those with $\rho_j^* < \hat{\rho}$ are redlined. Because banks order the borrower groups and because $\rho_G^* > \rho_j^*$, target groups are the residual borrowers. As a consequence, if a target group is rationed, the marginal effects of changes in supply or demand fall completely on the rationed group, while general borrowers are left unaffected. An interesting implication of this result occurs when one target group is rationed and another is redlined: the burden of subsidies to the previously redlined group falls entirely on the rationed group. In this case, credit subsidies simply rearrange loans among target groups; aggregate targeted borrowing does not rise.

A second important feature of the model is that the ordering of projects by borrower returns and social returns will differ, in gen-

¹⁵Redlining equilibria are identical to market-clearing equilibria, except that the index runs over only those groups with positive loan values.

¹⁶There is an infinitesimal chance that these groups clear or are redlined, depending on the position of S_2 in Figure 2.

TABLE 2—CHARACTERIZATION OF FEDERAL CREDIT POLICIES, 1980–1987

Sector	Principal credit instrument	s	s^G	s^O	γ	δ	σ
Mortgage	loan guarantees	0.02	< 0.01	0.02	1.0	1.0	0.01
Farm	direct lending	0.25	0.00	0.25	0.0	0.0	0.00
Student	loan guarantees	0.32	0.04	0.28	1.0	1.0	0.39
Small business	loan guarantees	0.14	0.09	0.00	0.9	0.0	0.91
Tax-exempt	interest subsidy	0.19	0.00	0.19	0.0	0.0	0.00
Variable	Definition						
s	overall reduction in the present discounted value of loan payments induced by federal credit subsidies $= s^O + s^G$						
s^G	reduction in the present discounted value of loan payments induced by just the subsidized guarantee						
s^O	reduction in the present discounted value of loan payments due to other factors						
γ	proportion of principal and interest covered by the guarantee						
δ	proportion of guarantee fee paid by the borrower						
σ	subsidized portion of the guarantee fee						

Sources: See text and Appendix B.

eral. As a consequence, subsidies can raise or reduce welfare (defined as total surplus), depending on the investment response and the size of the subsidy (Gale, 1990b). Whether actual policies raise welfare is examined below.

III. Parameter Specification

The specification of federal credit programs is potentially complicated due to the large number and diversity of credit programs, as discussed in Section I. However, a sizeable percentage of both the credit volume and the subsidies inherent in credit programs is concentrated in a relatively small number of programs and sectors. In 1986, the five sets of programs listed in Table 2 accounted for 93 percent of all new credit advanced, subsidized, or guaranteed by the government, including 75 percent of direct loan obligations and 90 percent of guaranteed loans.^{17,18} The subsidies in these

programs accounted for approximately 75 percent of all subsidies extended through loan guarantees and 50 percent of subsidies extended through direct lending (excluding foreign military sales credit). In order to capture the main effects of credit interventions without needlessly complicating the analysis, the simulations focus on these programs. Table 2 specifies their underlying parameters. A brief discussion is presented here; detailed references are given in Appendix B.

The overall percentage reduction in the present discounted value of a borrower's loan payments induced by a federal subsidy is given by s and is estimated by the Office of Management and Budget.¹⁹ All target groups except mortgage borrowers receive substantial subsidies from the government. These benefits may appear in one of two forms: a subsidized fee for a loan guarantee (s^G) or "other" forms (s^O), including reduced interest payments, deferral of payments, grace periods, and longer maturities. A loan-guarantee fee is considered to be subsidized if it does not cover the expected

¹⁷These figures omit government-sponsored enterprises (GSE's) other than the Farm Credit System, which are also omitted from the simulations, in order to avoid the need to model the operation of secondary markets. However, since GSE's are concentrated in mortgage, farm, and student credit markets, the importance of credit subsidies to the five sectors listed would increase if GSE's were included.

¹⁸The programs listed in Table 2 are in some cases meant to refer to combinations of individual programs, as discussed in Appendix B.

¹⁹Except for tax-exempt borrowers, estimates may be found annually in "Special Analysis F," *Special Analyses: Budget of the United States*. Although the estimates are necessarily rough, they serve to provide a general impression of the overall subsidy rate. The calculation of the overall subsidy for tax-exempt borrowers is provided in Appendix B.

TABLE 3—SPECIFICATION OF OTHER PARAMETERS

Sector	L	β	ϕ	r^*	\hat{r}
Mortgage	4.6	-1.80	0.97	0.101	0.103
Farm	2.0	-1.00	0.88	0.078	0.118
Student	0.6	-0.65	0.91	0.032	0.122
Small Business	0.2	-0.80	0.90	0.092	0.127
Tax-Exempt	8.3	-0.40	1.00	0.054	0.099
General Borrowers	81.9	-0.80	1.00	0.100	0.099

Variable	Definition
L	credit allocation as percentage of total credit
β	interest elasticity of loan demand
ϕ	repayment rate
r^*	effective interest rate
\hat{r}	interest rate in the absence of credit subsidies

Notes: Private credit allocations (L) sum to only 97.6, rather than 100, because government for credit subsidies requires 2.4 percent of available credit. The parameters r^* and \hat{r} were calculated on the assumption that $\rho = 0.10$.

Source: See text and Appendix B.

default costs. The composition of the overall subsidy varies considerably across programs.

The parameters used to calculate s^G and s^O are listed next: γ is the guarantee rate, and δ is the proportion of the fee paid by the borrower; both are obtained from government documents. The percentage of default costs not covered by the guarantee fee is given by σ and is calculated in Appendix B.

Table 3 presents estimates of other key parameters. Credit allocations are normalized to equal 100 and are based on recent historical averages. The allocations for each target group are meant to include only federal lending. Consequently, unsubsidized loans (including those for mortgages, farms, etc.) are included in general borrowing. Demand elasticities are obtained from recent studies.

Repayment rates are specified as constants, rather than functions of interest rates (as in Section II), for simplicity and due to the lack of any useful data on the shape of the function. It is still possible to generate rationing outcomes by specifying a maximum effective interest rate for each group and assuming that banks will not lend to the target group at higher rates. The effect of this assumption is to generate a ρ - r curve

that rises linearly to an interior maximum at r_j^* and then falls. Thus, the specification of constant ϕ 's considerably simplifies the calculations without losing any important aspect of the underlying model. For general borrowers, ϕ is set to 1.0. Other reported repayment rates are based on government data and existing studies.

Effective interest rates (r^*) are calculated (as described in Appendix A) on the assumption of $\rho = 0.10$. Estimated interest rates in the absence of credit subsidies are presented in the last column. Credit subsidies lower effective target-group interest rates below those paid by general borrowers for all target groups except mortgage borrowers.

Perhaps the single most important and controversial parameter is the elasticity of supply of funds, or more generally the openness of capital markets. A large literature on this topic has developed from papers by Arnold Harberger (1978, 1980) and Martin Feldstein and Charles Horioka (1980), with mixed conclusions. The supply of credit from domestic sources is also controversial. Robert DeFina (1984) surveys existing results and lists interest elasticities ranging from 0 to 5. This paper employs two alternative assumptions concerning the elasticity of supply: $\alpha = 5.0$ and $\alpha = 0.5$. Using

these extremes will provide bounds for the effects of credit policies and will help indicate which results are robust to alternative views about capital markets.

IV. Results

Before discussing the effects of credit policies, it is necessary to resolve two further issues: 1) the interpretation of existing credit allocations (summarized in Table 3) as representing market-clearing or rationed outcomes and 2) assumptions concerning the allocation of credit without any lending subsidies. I specify two interpretations of existing credit patterns and policies: one in which all groups clear and one in which farmers are rationed but other groups clear.²⁰ The allocation of credit without lending subsidies depends on assumptions about whether the market is rationed with existing credit policies, as well as assumptions concerning the shape of the $p-r$ curve for each group. If all markets clear with existing policies, I examine two scenarios in the world without credit subsidies: (a) all markets clear and (b) tax-exempt and mortgage markets clear, but farmers, students, and small businesses are redlined. These cases are meant to represent optimistic and pessimistic extremes concerning the viability of target-group markets in the absence of credit policy. If the existing credit allocation represents rationing of farmers, I assume that case (b) holds without credit policy.

A. Allocation and Level of Credit

The effects of credit subsidies on the allocation and level of credit are presented in Table 4. Starting from a world with no credit subsidies, as represented by scenario (a) or (b) described above, the introduction of all credit programs generates the base-case allocations listed in Table 3 (reproduced here

for convenience). Many of the aggregate effects depend on the elasticity of supply.

In the high-elasticity case ($\alpha = 5$), credit programs raise private borrowing by between 2.5 percent [scenario (b)] and 4.1 percent [scenario (a)]. Crowding-out effects are very small: general borrowing falls by approximately 1 percent in either scenario. Aggregate target-group investment increases by between 23 percent and 42 percent. With the exception of mortgage borrowers, all target groups experience substantial increases in investment.

If the supply of funds is inelastic ($\alpha = 0.5$), different aggregate patterns emerge. Credit programs have essentially no net effect on aggregate private borrowing. General borrowing falls by between 3.5 percent in scenario (a) and 4.6 percent in scenario (b) but is matched by an equivalent absolute rise in targeted lending.

Table 4 indicates that credit programs exert powerful effects on the allocation of credit to target groups, regardless of the supply elasticity. Other simulations (not reported) indicate that these sectoral effects remain for a wide range of demand elasticities as well.

B. Welfare Effects

Standard welfare analysis (see Harberger, 1971) aggregates consumers' and producers' surplus and assumes that the government has access to lump-sum taxes. The welfare analysis described here is based on the traditional approach, with two qualifications. First, consumers' surplus is multiplied by the probability of repayment, so that borrowers obtain the surplus only if their project succeeds. Second, loan demand is assumed to become flat at $r^* = 50$ percent and to be zero above that point. This affects the level of consumers' surplus but will not affect the difference in welfare effects across policies, unless a group is redlined.

The simulations indicate that credit programs generate sizable efficiency costs. Using 1980–1987 averages for net credit and gross national product (GNP), credit programs are estimated to reduce welfare by 0.27–0.40 percent of GNP. Using 1987 val-

²⁰I assume that farmers are rationed at 80 percent of their desired quantity of loans. It is straightforward to analyze the effects of rationing on other groups as well. The important analytical point, however, is how effects differ across rationing and clearing regimes, which may be illustrated with any single group.

TABLE 4—ALLOCATION OF FUNDS UNDER ALTERNATIVE SCENARIOS

Sector	Base case	$\alpha = 5.0$		$\alpha = 0.5$	
		(a)	(b)	(a)	(b)
Mortgage	4.60	4.48	4.50	4.75	4.87
Farm	2.00	1.31	0	1.36	0
Student	0.60	0.25	0	0.26	0
Small business	0.20	0.15	0	0.16	0
Tax-exempt	8.30	6.50	6.51	6.58	6.62
Total, target groups	15.70	12.69	11.01	13.10	11.49
General borrowers	81.91	82.55	82.76	84.78	85.72
Total, all groups	97.61	95.24	93.77	97.88	97.21
ρ	0.1000	0.0990	0.0987	0.0958	0.0945

Notes: Scenarios are described in the text. For the base case, government borrowing to fund credit programs = 2.39, so total credit sums to 100; for the alternative cases (a) and (b), government borrowing is zero.

TABLE 5—GOVERNMENT COSTS (\$) PER DOLLAR OF INCREMENTAL TARGETED INVESTMENT

Sector	Demand elasticities					
	High		Original		Low	
	(a)	(b)	(a)	(b)	(a)	(b)
Mortgage	0.54	0.66	0.74	0.92	1.29	1.70
Farm	0.53	0.25	0.72	0.25	1.31	0.25
Student	0.43	0.32	0.55	0.32	0.91	0.32
Small business	0.44	0.14	0.62	0.14	1.16	0.14
Tax-exempt	0.62	0.62	0.87	0.88	1.64	1.65
Overall:	0.57	0.43	0.79	0.51	1.44	0.63

Notes: All of the estimates assume that $\alpha = 5.0$. Original demand elasticities are given in Table 3; high (low) elasticities raise (reduce) the original figures by 50 percent.

ues, the welfare loss is 0.25–0.37 percent of GNP, or approximately \$10–15 billion.²¹

The estimates vary somewhat due to assumptions concerning supply and demand elasticities and initial scenarios. Nevertheless, several aspects of the results warrant comment. First is the sheer magnitude: at

1/3 percent of GNP, the welfare loss due to credit programs is substantial. Second, since the results are fairly insensitive to specification of supply or demand elasticities, it is not possible to infer the efficiency costs from the allocational effects. In particular, the absence of crowding out implies only that the costs are borne by agents other than nontargeted investors.

Third, the results indicate the importance of explicitly considering the financing costs of credit programs. For most simulations, the gains and losses to the various borrower groups and suppliers of funds roughly offset each other. *The welfare loss occurs because the programs must be financed.* As shown in Table 5, government costs per dollar of incremental targeted investment are ex-

²¹ These numbers are generated by multiplying the welfare changes when the total credit allocation is 100 by (actual net credit)/100 and dividing by GNP. If demand curves do not become flat until $r^* = 100$ percent, the efficiency cost falls by approximately 5 percent. If resource costs of distortionary taxation are included at 30 cents per dollar of government revenue (see Charles Ballard et al., 1985; Kenneth Judd, 1987), the efficiency cost rises to 0.37–0.55 percent of GNP.

traordinarily high.²² Using $\alpha = 5$ and the original demand elasticities, the costs exceed 50 cents for each group when starting from scenario (a). For scenario (b), the costs are somewhat lower because several groups are redlined initially, but the average cost is still 50 cents. These results are not sensitive to shifts in α . However, increasing demand elasticities by 50 percent raises the average cost to over 90 cents for any group that is not redlined beforehand.

In a perfectly efficient world, the traditional approach would provide all of the necessary information for welfare analysis. However, focusing solely on the aggregate welfare effects may be misleading because, as discussed above, credit programs are often enacted specifically in response to some preexisting distortion, externality, or distributional concern. Although it is beyond the scope of this paper to develop a model explicitly incorporating all of these issues, several useful insights concerning the welfare effects of credit policy can be derived.

First, under all scenarios, all target groups except mortgage borrowers experience large gains in consumer surplus. However, because the programs subsidize both inframarginal and incremental borrowers and because there are no societal benefits to subsidizing the former group, the division of direct welfare benefits should be an important policy concern.

Under scenario (a), the vast majority of direct welfare benefits accrue to target-group members who would have received credit without public assistance. Approximately 76 percent of student welfare gains are inframarginal. For farmers, small businesses, and tax-exempts, 90 percent of the benefits constitute lump-sum transfers. More than 99 percent of mortgage borrowers' welfare gains also fall into this category. Under scenario (b), of course, none of the

gains made by farmers, students, or small businesses are inframarginal.

Therefore, whether credit programs principally aid inframarginal borrowers or new borrowers depends on assumptions concerning the status of sectoral credit markets without lending subsidies. Because credit policy has been an integral part of domestic credit markets for many years, it is difficult to determine with precision which scenario, (a) or (b), is closer to reality.²³ However, whenever the target market clears, any *marginal* increase in subsidies will benefit primarily inframarginal borrowers.

A second issue is the existence of external benefits. Although specification and calculation of external benefits would be extremely complicated, it is possible to calculate how large the benefits would need to be to offset the welfare losses described above. For each program, it is convenient to express the minimum level of external benefits needed to raise welfare as a percentage of the change in the target group's investment. These percentages are called threshold levels and are shown in Table 6.²⁴

Sizable external benefits must occur for current policies to raise welfare. Using the original demand elasticities and $\alpha = 5$, external benefits must be 40–70 percent of the change in target-group investment, if the target-group market clears beforehand. Thresholds for marginal changes in policy are slightly lower than for overall changes. Threshold levels for redlined groups are considerably lower, because the change in target-group investment is much larger. For

²²These figures are calculations of the model, given the definition of government borrowing in equation (6). The numbers bear no relation to any reported budget figure. Federal budgetary treatment of credit programs is highly misleading and inconsistent (see Bosworth et al. [1987], the Congressional Budget Office [1984] or Gale [1990a] for discussion).

²³It is not obvious that students (who have no collateral or credit history) and small businesses (who must be turned down by at least two banks to qualify for a guaranteed loan) would be redlined in the absence of public assistance. For example, parents could borrow on behalf of their children. SBA guarantees are typically for larger loans with longer maturities than private credit to small business. Without guarantees, borrowers could perhaps qualify for shorter-term or smaller loans. In addition, under current SBA regulations, lenders have incentives to reject private applications in hope of acquiring a guarantee.

²⁴Since the estimates assume that investment crowded out by a credit policy contains no external benefits, the threshold levels are lower bounds on the level of external benefits required to raise welfare.

TABLE 6—THRESHOLD LEVELS (PERCENTAGE) FOR ALTERNATIVE CREDIT POLICIES

Sector	Target-group market with no credit subsidy	Proposed policy	Demand elasticity		
			High	Original	Low
Mortgage	clearing	current policy	30	45	90
	clearing	increase in subsidy	31	46	90
Farm	clearing	current policy	52	70	125
	redlined	current policy	13	14	41
Student	clearing	increase in subsidy	55	68	109
	clearing	current policy	41	51	129
	redlined	current policy	26	21	11
	clearing	increase in subsidy	43	47	60
Small business	clearing	current policy	38	57	93
	redlined	current policy	3	~ 0 ^a	-9
Tax-exempts	clearing	increase in subsidy	40	52	89
	clearing	current policy	55	77	143
	clearing	increase in subsidy	50	66	108

Notes: Threshold levels are defined in the text. Estimates use $\alpha = 5.0$. Original demand elasticities are provided in Table 3; high (low) elasticities raise (reduce) the original estimates by 50 percent.

^aBetween 0 and -0.1.

farmers and students, threshold levels fall to 14 percent and 21 percent, respectively. Although these figures may appear to be encouragingly low, they are noteworthy primarily because they are still positive: even if the government creates the market for farm or student loans (with current policies), there is a welfare loss unless sufficient *external* benefits exist. In contrast, SBA guarantees may generate welfare gains even in the absence of external benefits.²⁵

Threshold levels are not sensitive to shifts in α but rise dramatically for lower demand elasticities. Even in the most favorable case of high demand elasticities, threshold levels are on the order of 30–50 percent. Credit programs may generate external benefits of the appropriate magnitude to raise welfare. However, in the absence of any compelling empirical evidence that such benefits occur, the results serve primarily to emphasize the efficiency losses associated with credit interventions.

C. Interactions and the Effects of Credit Rationing

Because credit subsidies reallocate funds and raise the overall rate of return, a program that subsidizes one target group will

²⁵These results do not depend on the assumption that demand becomes flat at $r = 50$ percent.

necessarily affect credit allocations to other groups. The magnitude of the consequent crowding out is clearly an important policy concern. If policies completely or largely offset each other, then the government is spending substantial resources for little net gain.²⁶

Assume that, in the absence of credit programs, target group j receives $D_j(0)$ in funds. When only a subsidy for group j is introduced, the group receives $D_j(j)$. When all current programs exist, group j receives $D_j(\text{all})$. The offset is then measured by:

$$\text{OFFSET}_j = \frac{D_j(j) - D_j(\text{all})}{D_j(j) - D_j(0)}$$

That is, the offset is the percentage of the gain induced by the original credit subsidy that is eliminated by the introduction of all other credit programs. Estimates of offset percentages are shown in Table 7.

Starting from scenario (a) and setting $\alpha = 5$ yields very small interactions for all credit programs except mortgages. Almost two-fifths of the original rise in mortgage credit due to subsidies is eliminated by other

²⁶The importance of such interactions has been recognized independently by Bosworth et al. (1987). They do not quantify these effects, however. Penner and Silbert (1973) focus on interactions among programs that focus on the same target group.

TABLE 7—OFFSETS: THE EFFECTS OF INTERACTIONS AMONG PROGRAMS

Sector	Scenarios with no subsidy	Interpretation of base case	Offsets (percentage)	
			$\alpha = 5.0$	$\alpha = 0.5$
Mortgage	(a)	clearing	38.3	175.0
Farm	(a)	clearing	2.5	10.9
	(b)	rationed ^a	33.8	35.3
	(b)	rationed ^b	17.4	19.8
Student	(a)	clearing	2.6	11.0
Small business	(a)	clearing	3.5	14.9
Tax-exempt	(a)	clearing	1.2	4.9

Notes: Offsets are defined in the text. Estimates shown are based on the demand elasticities reported in Table 3.

^aRationed at 67 percent of notational demand.

^bRationed at 83 percent of notational demand.

programs. This large crowding-out effect occurs because the original subsidy to mortgages is small ($s = 0.02$). Thus, even a small increase in ρ eliminates a substantial percentage of the original benefits. For the other groups, the offset is small because direct subsidies are large.

At the lower value for α , interactions are much larger. For groups other than mortgage borrowers, the interactive effect quadruples: 15 percent of small-business investment gains and 10 percent of increases in farm and student credit are eliminated by other federal credit programs. Although these interactions are unlikely to have important macroeconomic effects, their effects on individual sectors are nonetheless sizable. For mortgage borrowers, the results are even more striking. They would be better off if all subsidies were eliminated. The crowding out induced by other subsidies more than offsets the small direct subsidy provided by government. Thus, as shown in Table 4, mortgage credit falls with the creation of all programs.

Interactions become extremely important when a group (e.g., farmers in Table 7) is rationed given current credit policies. If only a farm policy is introduced from scenario (b), farm credit rises to a point on farmers' notional demand curve. However, as other credit programs are introduced, interest rates rise and make the farmers' rationing constraint bind. Farmers' notional demand at that rate is typically much higher than the residual amount allotted to them. If farmers are rationed at 67 percent (83 percent) of

their notional demand, the offset is approximately 34 percent (18 percent). In any case, if a group is rationed in equilibrium, the offsets will typically be large, regardless of α . Offsets are not sensitive to variations in demand elasticities.

D. Policy Reform

This section uses the model to examine the implications of some recent events and proposals to reform federal credit. Because specific features of individual credit markets may be important in this regard, the results are necessarily exploratory.

Due to the high and rising costs of federal farm credit, many proposals have focused on reducing the subsidy rate. Halving the rate would reduce farm borrowing by 22 percent and would reduce government costs by 60 percent. Elimination would reduce farm credit by 35 percent.

Sharp increases in observed defaults at the FHA have directed a considerable amount of attention to modifying mortgage loan guarantees. If the overall default rate on FHA guarantees ultimately rises to 10 percent (from the 3 percent used above to represent earlier years), the welfare costs of mortgage guarantees would rise fourfold.²⁷ However, mortgage guarantees would still be the *least* subsidized of the five major credit programs analyzed here.

²⁷The results presented in this section are based on $\alpha = 5$, but are not sensitive to changes in α .

Although they occupy only a small portion of modern credit markets, guaranteed loans to students and small businesses have generated a substantial amount of controversy, presumably because of their perceived high costs. However, because of the differing structure of the programs, different methods for reducing costs are warranted. For example, as shown in Table 2, the principal subsidy to students arises from the stipulation that borrowers may defer principal and interest payments until nine months after graduation with no penalty. Elimination of that subsidy reduces the government's costs by 96 percent and reduces loan demand by 60 percent. However, most of the policy debate has focused on collecting defaulted student loans. Eliminating government costs from default (i.e., setting the guarantee fee equal to the loss rate) would reduce government costs by only 23 percent and would reduce demand by 14 percent.

In contrast, most of the benefits of SBA loan guarantees reside in the subsidized guarantee fee. Setting the guarantee fee equal to the loss rate would reduce small business borrowing by 16 percent, but government costs would fall by 70 percent. Thus, defaults are an important issue for SBA loan guarantees, whereas for students, program costs will continue to be large as long as deferral until graduation is allowed.

In its Preferred Lenders' Program, the SBA offers lending institutions significant reductions in paperwork and allows for greater lender discretion in making loans in exchange for reducing the guarantee rate to 75 percent from 90 percent. If the program works as intended, the increased flexibility will attract more lenders to the program, while the reduced guarantee will induce more careful screening. The effect of the program should be to reduce loss rates. If loss rates fall by 2.5 percentage points (from 10 percent to 7.5 percent), the allocation of funds to small business would fall by 1.5 percent, but small business welfare would *rise*, because the probability of repayment increases. Government borrowing for small business would fall by 30 percent and aggre-

gate welfare would *rise* slightly.²⁸ Although this result is speculative, the analysis indicates the potential for credit programs to raise welfare by improving the incentives and organization of sectoral credit markets, while still preserving the basic purpose and nature of the credit policy.

V. Conclusions

The role of government in imperfect capital markets is an issue of obvious theoretical and practical concern. This paper has developed a model of the credit market and simulated the effects of federal credit interventions. Major results are reported in the introduction. Some caveats and possible extensions should be mentioned at this stage. First, because any simulation employs a variety of maintained assumptions and approximations, results should be interpreted as showing the direction and relative magnitude of various effects, rather than as precise estimates.

Second, since the model assumes that all agents are risk-neutral, all assets are perfect substitutes, and rates of return are equalized across assets. However, many assets may not be good substitutes in asset demand (Benjamin Friedman, 1978; Jeffrey Frankel, 1985). In the present context, credit for students, small businesses, and farmers may be poor substitutes for mortgages, tax-exempts, or corporate debt. In that case, if lenders are risk-averse, the crowding-out effect of subsidies aimed at the first set of target groups would be smaller than indicated above, but so would the direct benefits (Penner and Silber, 1973; Friedman, 1978). With a range of substitutabilities among assets, credit policies could actually crowd in investment by other groups (Friedman, 1978). However, given that 75 percent of all net credit advanced is funneled through financial intermediaries (Michael Moran, 1985), which presumably face competitive pressures to maximize profits, the

²⁸Although screening costs would increase, they would at least partially be offset by reduced lender paperwork for the SBA.

assumption of risk neutrality may be a good approximation.²⁹

Sensitivity analysis on a broader scale would consider using an alternative theoretical framework. For example, Stephen Williamson (1987) and de Meza and Webb (1988) rely on monitoring or screening costs to generate imperfect markets. However, simulations of such models may prove difficult, due to the absence of any reliable estimates of those costs. Although de Meza and Webb (1987) show that allowing equity finance can fundamentally change the results derived from the Stiglitz-Weiss framework, borrowers served by federal credit typically do not issue external equity, so eliminating equity contracts does not appear to be a major restriction in this case.

In light of these results, several features of the model presented above make it an excellent candidate for simulations: 1) it allows comparison of market-clearing, rationing, or redlining outcomes; 2) results depend on parameters for which a fair amount of empirical information is available, rather than on screening or monitoring costs; 3) it easily accommodates multiple borrower groups; and 4) welfare effects depend on the particular parameter values, rather than being predetermined by theoretical factors.

Even if alternative models were used, my conjecture is that the basic results would stand: credit policy has important allocational effects, due to the large subsidies provided, and sizable efficiency costs, due to high government-financing costs. These aspects of credit policy appear to drive the results and would have to be incorporated in any theoretical structure.

Perhaps the most important and most difficult extension concerns the connection between credit allocations and real economic activity. The model maintains a simple, direct link: an increase in credit leads to increased real activity. D. C. Rao and Ira Kaminow (1975), Mary Kay Plantés and David Small (1981), and Bosworth et al.

(1987) emphasize that subsidized credit may simply induce borrowers to substitute debt for equity or capital for labor. The borrower may also use the funds for some entirely unrelated purpose. Under this view, changes in credit may not induce any shift in real activity.

Alternatively, if it provided the marginal source of funds for a large project, government may raise private investment by more than the government credit extended. In addition, a subsidy may serve to keep an enterprise viable now and raise investment in the future. Under these scenarios, investment rises by more than the change in government credit.

This paper makes no attempt to resolve these issues. Instead, the one-to-one link between credit and investment may be taken as a benchmark. Future research on federal credit may help to clarify real-financial linkages, perhaps through empirical analyses of individual programs. Another issue is how the design of credit programs influences their effects. It is easy to show that guarantees and subsidies generate differing incentive effects (Gale, 1987). Focusing on incentives can explain some of the divergent default behavior across programs, which were simply taken as given above.

Credit programs may also help explain long-term declines in saving. Student and mortgage guarantees reduce the level of initial wealth required to make sizable purchases of education and housing and thereby reduce one motive to save. Fumio Hayashi et al. (1987) find limited support for this proposition in comparing U.S. and Japanese saving data. Finally, it may be possible to exploit changes in credit programs to test for the nature and importance of liquidity constraints.

APPENDIX A: MODEL CALCULATIONS

Effective Interest Rates

Consider a \$1 loan and assume that all defaults occur in the first period of repayment. Given the probability of repayment, ϕ , the costs of funds, ρ , and the maturity period for loans, m , banks require constant

²⁹However, given the presence of federal deposit insurance, banks may be risk-seeking.

annual loan payments, X_1 , such that

$$(A1) \quad X_1 = \left[\phi \sum_{k=1}^m (1+\rho)^{-k} \right]^{-1}.$$

The implied interest rate, r_1 , is defined by

$$(A2) \quad X_1 = \left[\sum_{k=1}^m (1+r)^{-k} \right]^{-1}.$$

Substitution of (A1) into (A2) yields

$$(A3) \quad \sum_{k=1}^m [(1+\rho)^{-k}] \\ = \phi^{-1} \sum_{k=1}^m [(1+r_1)^{-k}].$$

Equation (A3) describes the determination of effective interest rates, given ρ , ϕ , and m , in the absence of credit subsidies. If the government introduces a loan guarantee, the bank now requires annual payments, X_2 , such that its discounted costs equal its discounted expected receipts, or

$$(A4) \quad \phi X_2 \sum_{k=1}^m [(1+\rho)^{-k}] + (1-\phi)\gamma \\ = 1 + (1-\delta)(1-\sigma)(1-\phi)\gamma.$$

The left side represents expected bank revenues. The last term on the right represents bank payments for the guarantee fee: the fair-insurance cost of the guarantee is $(1-\phi)\gamma$, but the government pays σ , and borrowers pay δ .

The percentage reduction in the present value of loan payments induced by the subsidized loan guarantee, s_G , is found by solving

$$(A5) \quad 1 - s_G \\ = \frac{\delta(1-\sigma)(1-\phi)\gamma + \sum_{k=1}^m X_2/[(1+r_1)^{-k}]}{\sum_{k=1}^m X_1/[(1+r_1)^{-k}]}.$$

This expression gives the ratio of borrower

payments with and without the guarantee. Using (A2) and some algebra, (A5) reduces to

$$(A6) \quad s_G = \sigma(1-\phi)\gamma.$$

Define X_3 as annual borrower payments with guarantees and other subsidies, and define s as the percentage reduction in the present value of loan payments due to all aspects of credit policy. Then,

$$(A7) \quad 1 - s \\ = \frac{\delta(1-\sigma)(1-\phi)\gamma + \sum_{k=1}^m \{X_3/[(1+r_1)^{-k}]\}}{\sum_{k=1}^m \{X_1/[(1+r_1)^{-k}]\}}.$$

Calculations similar to those used in deriving (A6) yield

$$(A8) \quad X_3 = X_1[1 - s - \delta(1-\sigma)(1-\phi)\gamma].$$

Other subsidies ($s^O = s - s_G$) are modeled as further reductions in annual loan payments. Substitution yields

$$(A9) \quad X_2 = X_3 + \frac{s^O}{\phi \sum_{k=1}^m \frac{1}{(1+\rho)^k}}.$$

The effective interest rate for borrowers, r^* , solves

$$(A10) \quad X_3 = \frac{1 - \delta(1-\sigma)(1-\phi)\gamma}{\sum_{k=1}^m [(1+r^*)^{-k}]}.$$

Substitution of (A10) into (A9) and rearranging yields

$$(A11) \quad \frac{1 - [1 - (1-\delta)(1-\sigma)](1-\phi)\gamma - s^O}{\phi[1 - \delta(1-\sigma)(1-\phi)\gamma]} \\ \times \sum_{k=1}^m \frac{1}{(1+r^*)^k} = \sum_{k=1}^m \frac{1}{(1+\rho)^k}.$$

Equation (A11) implicitly defines r^* . In the absence of subsidies, (A11) reduces to (A3).

Government Borrowing

For a guarantee, the government receives $(1-\sigma)(1-\phi)\gamma$ in guarantee fees but expects to pay $(1-\phi)\gamma$ in defaults. Thus, the government cost of a guarantee is s_G as defined by (A6). In addition, when providing the other subsidies described above, the government pays $X_2 - X_3$ with probability ϕ for each period of the loan. The cost to government in each period is $\phi(X_2 - X_3) = \phi s^O X_1 = \phi(s - s_G)X_1$. Total government borrowing requirements, discounted at ρ , are

$$(A12) \quad G_{LG} = s_G + (s - s_G)\phi X_1 \\ \times \sum_{k=1}^m \frac{1}{(1+\rho)^k} = s.$$

Under a direct loan program, receipts are X_3 per period with probability ϕ , and 0 with probability $1-\phi$. From (A8) and (A1), $X_3 = X_2 - s^O X_1$; but since $X_2 = X_1$ for a direct loan, $X_3 = (1-s^O)X_1$. Therefore, expected government borrowing is given by

$$(A13) \quad G_{DL} = 1 - \phi(1-s^O) \sum_{k=1}^m \frac{X_1}{(1+\rho)^k} \\ = s^O = s.$$

In a tax-exempt program, the government forgoes $X_2 - X_3$ in receipts each period with probability ϕ . Government financing costs are given by

$$(A14) \quad G_{TE} = \phi \sum_{k=1}^m \frac{X_2 - X_3}{(1+\rho)^k} \\ = s^O = s.$$

A more detailed version of this Appendix is available upon request.

APPENDIX B: PARAMETER SPECIFICATION

Unless otherwise specified, data are collected from annual versions of "Special Analysis F" of *Special Analyses: Budget of*

the U.S. Government. Primary mortgage guarantees issued by the Federal Housing Administration (FHA) and the Veteran's Administration (VA) cover 100 percent of the outstanding principal and interest ($\gamma = 1$) and are paid for by the borrower ($\delta = 1$). Default rates of 3.0 percent were assumed for both programs, based on evidence from the mid-1980's. FHA fees are 3.8 percent; VA fees are 1.9 percent. Weighting by new credit issued yields the value for σ . The overall reduction in borrower payments is estimated to be 1 percent for FHA and 4.2 percent for VA guarantees. Weighting by new credit issued yields an average reduction of 2 percent.

Student loans are 100-percent guaranteed. Loss rates due to default are approximately 9 percent (Bosworth et al., 1987 pp. 134-5). Therefore, $\phi = 0.91$. Students currently pay the full fee of 5.5 percent, implying $\delta = 1$ and $\sigma = 0.39 (= 1 - 0.055/0.09)$. Other characteristics of the loan program, particularly the deferral of principal and interest at no cost until graduation, substantially reduce borrower costs. Students would have had to pay 41 percent, 47 percent, and 52 percent more (in present-value terms) for loans in 1984, 1985, and 1986, if there were no guarantee program (table F-12 in *Special Analyses: Budget of the U.S. Government*). Therefore, the guarantee reduced borrower payments (in present value) by 29 percent, 32 percent, and 34 percent respectively, in those years. The value of 32 percent is chosen to represent s .

Small Business Administration (SBA) loan guarantees provide timely payment of 90 percent of outstanding principal and interest. The loss rate on SBA guarantees has been estimated at 9.7 percent (Bosworth et al., 1987 p. 93) and 12.5 percent (Michael Boskin and Bradford Barham, 1984 p. 28). Since fees in the mid-1980's were 1 percent of loan volume and are paid by the bank ($\delta = 0$), a loss rate of 11 percent ($\phi = 0.89$) implies that $\sigma = 0.91$. During 1984-1986, SBA guarantees reduced the present value of borrower payments by 10.8 percent, 14.1 percent, and 18.6 percent (based on data in table F-12). Therefore, s is set equal to 0.14.

For tax-exempt borrowing, the only parameter to set is s , the overall reduction in borrower payments. If the non-tax-exempt borrowing rate is 10 percent (the value of ρ below), and tax-exempt rates are 7 percent (a spread consistent with results reported in James Poterba [1986]), tax-exempt status reduces the discounted (at 10 percent) value of borrower payments on a 10-year loan by 19 percent. The rate of default for tax-exempts is set equal to zero.

Direct farm lending is meant to represent an amalgam of government lending programs to the agricultural sector. These include programs in the Farmers' Home Administration (FmHA), Rural Electrical Administration, and Export Credits. During 1984–1986, the loan programs reduced borrower payments by 21.4 percent, 20.0 percent, and 30.0 percent (table F-11). A rough average is $s = 0.25$. Calomiris et al. (1986 table A-2) report delinquency rates of approximately 4.0 percent of loan volume for all agricultural credit during 1982–1985. However, federal credit programs in agriculture undoubtedly have higher default rates; the FmHA is estimated to have held 4.5 percent of its portfolio in delinquent loans in 1980, rising to 23.0 percent in 1985 (General Accounting Office, 1986 p. 53). A loss rate of 0.12 ($\phi = 0.88$) is used in the simulation.

The demand elasticity for general borrowers is assumed to be -0.8 , based on estimates by Robert Hall (1977), Friedman (1978), and Feldstein and Joosung Jun (1987). In the absence of better data, this value is also used for small businesses. The elasticity of mortgage borrowing is set at -1.8 , based on Phoebe Dhyrnes and Paul Taubman (1969). Farm-investment elasticity is set at -1.0 , based on Michael LeBlanc and James Hrubovcak (1986). The elasticity of student loan demand is estimated from figures in Bosworth et al. (1987). The tax-exempt demand elasticity is based on Patric Hendershott and Timothy Koch (1977). Loan maturities were set at 30 years for mortgages, 10 years for students, small businesses, and tax-exempts, and 20 years for farmers, based on data in table F-11 for 1985 and 1986.

Credit allocations are based on flow-of-funds figures from recent years. Since these data are based on net borrowing, while credit subsidies accompany new credit, some adjustments are necessary. Tax-exempt borrowing represented 8.3 percent of net borrowing during 1980–1987 (see Table 1). FHA and VA mortgages accounted for 4.6 percent of net borrowing since 1980. Agricultural lending represented less than 0.1 percent of net borrowing during 1980–1987, because net agricultural lending was negative in several years. Nonetheless, many new agricultural loans were made. During 1970–1985 and 1976–1980, farm credit averaged 4.1 percent and 4.2 percent of net credit, respectively. Since the government has typically provided between one-third and one-half of all agricultural lending (Council of Economic Advisers, 1986 p. 197), federal farm credit is set at 2 percent of aggregate credit. Net guaranteed student loans for 1980–1987 totalled \$29.4 billion, or 0.6 percent of total net borrowing. Net SBA loan guarantees have been negative in recent years, but new SBA disbursements were approximately one-third of guaranteed student lending during 1980–1987, so the SBA allocation is set at 0.2 percent.

Finally, ρ is set at 0.1. Other rates are calculated given equation (A11). Given these parameters, the model may be solved backwards to generate values of a and b_i , the constants in the supply and demand equations. Alternative initial levels of ρ have no appreciable effects on the results, because a and b_i adjust accordingly.

REFERENCES

- Ballard, Charles L., Shoven, John B. and Whalley, John, "General Equilibrium Computations of the Marginal Welfare Costs of Taxes in the United States," *American Economic Review*, March 1985, 75, 128–39.
- Boskin, Michael J. and Barham, Bradford L., "Measurement and Conceptual Issues in Federal Budget Treatment of Loans and Loan Guarantees," CEPR Publication No. 11, Stanford University, 1984.

- Bosworth, Barry P., Carron, Andrew S. and Rhyne, Elizabeth, *The Economics of Government Credit*, Washington, DC: Brookings Institution, 1987.
- Calomiris, Charles W., Hubbard, R. Glenn and Stock, James H., "Growing in Debt: The 'Farm Crisis' and Public Policy," *Brookings Papers on Economic Activity*, 1986, (2), 441-79.
- Chaney, Paul K. and Thakor, Anjan V., "Incentive Effects of Benevolent Intervention: The Case of Loan Guarantees," *Journal of Public Economics*, March 1985, 26, 168-89.
- DeFina, Robert H., "The Link Between Savings and Interest Rates: A Key Element in the Tax Policy Debate," *Business Review*, Federal Reserve Bank of Philadelphia, November 1984, 15-21.
- de Meza, David and Webb, David C., "Too Much Investment: A Problem of Asymmetric Information," *Quarterly Journal of Economics*, May 1987, 101, 281-92.
- _____, and _____, "Credit Market Efficiency and Tax Policy in the Presence of Screening Costs," *Journal of Public Economics*, June 1988, 36, 1-22.
- Dhrymes, Phoebus J. and Taubman, Paul J., "An Empirical Analysis of the Savings and Loan Industry," in Irwin Friend (ed.), *Study of the Savings and Loan Industry*, Washington, DC: U.S. Government Printing Office, 1969, 67-181.
- Feldstein, Martin and Horioka, Charles, "Domestic Saving and International Capital Flows," *Economic Journal*, June 1980, 90, 314-29.
- _____, and Jun, Joosung, "The Effects of Tax Rules on Nonresidential Fixed Investment," in Martin Feldstein, ed., *The Effects of Taxation on Capital Accumulation*, Chicago: University of Chicago Press, 1987, 101-56.
- Friedman, Benjamin, "Crowding Out of Crowding In? Economic Consequences of Financing Government Deficits," *Brookings Papers on Economic Activity*, 1978, (3), 593-641.
- Frankel, Jeffrey A., "Portfolio Crowding Out, Empirically Estimated," *Quarterly Journal of Economics*, 1985 special supplement, 100, 1041-65.
- Gale, William G., "The Allocational and Welfare Effects of Federal Credit Programs," Ph.D. Dissertation, Stanford University, 1987.
- _____, "Collateral, Rationing, Government Intervention in Credit Markets," NBER Working Paper 3024, July 1989; (forthcoming in R. Glenn Hubbard, ed., *Asymmetric Information, Corporate Finance, and Investment*, Chicago: University of Chicago Press).
- _____, (1990a) "The Hybrid Plan: A Proposal for Federal Credit Reform," *Contemporary Policy Issues*, April 1990, 8, 107-21.
- _____, (1990b) "Federal Lending and the Market for Credit," *Journal of Public Economics*, July 1990, 42, 177-93.
- Hall, Robert E., "Investment, Interest Rates, and the Effects of Stabilization Policies," *Brookings Papers on Economic Activity*, 1977, (1), 61-103.
- Harberger, Arnold C., "Three Basic Postulates for Applied Welfare Economics," *Journal of Economic Literature*, June 1971, 9, 785-97.
- _____, "Perspectives on Capital and Technology in Less Developed Countries," in M. J. Artis and A. R. Nobay, eds., *Contemporary Economic Analysis*, London: Croom Helm, 1978, 42-72.
- _____, "Vignettes on the World Capital Market," *American Economic Review*, May 1980, 70, 331-7.
- Hardin, Clifford M. and Denzau, Arthur T., *The Unrestrained Growth of Federal Credit Programs*, Formal Publication No. 45, St. Louis, MO: Center for the Study of American Business, 1981.
- Hayashi, Fumio, Ito, Takatoshi and Slemrod, Joel B., "Housing Finance Imperfections and Private Saving: A Comparative Simulation Analysis of the United States and Japan," NBER Working Paper No. 2272, June 1987.
- Hendershott, Patric and Koch, Timothy, "An Empirical Analysis of the Market for Tax-Exempt Securities: Estimates and Forecasts," monograph 1977-4, Center for the Study of Financial Institutions, New York University, 1977.
- Judd, Kenneth L., "The Welfare Cost of Fac-

- tor Taxation in a Perfect Foresight Model," *Journal of Political Economy*, August 1987, 95, 675-709.
- LeBlanc, Michael and Hrubovcak, James**, "The Effects of Tax Policy on Aggregate Agricultural Investment," *American Journal of Agricultural Economics*, November 1986, 68, 767-77.
- Mankiw, N. Gregory**, "The Allocation of Credit and Financial Collapse," *Quarterly Journal of Economics*, August 1986, 101, 455-70.
- Moran, Michael J.**, "The Federally Sponsored Credit Agencies: An Overview," *Federal Reserve Bulletin*, June 1985, 71, 373-88.
- Penner, Rudolph G. and Silber, William L.**, "The Interaction Between Federal Credit Programs and the Impact on the Allocation of Credit," *American Economic Review*, December 1973, 63, 858-72.
- Plantes, Mary Kay and Small, David**, "Macroeconomic Consequences of Federal Credit Activity," in *Conference on the Economics of Federal Credit Activity*, Special Study, Washington, DC: Congressional Budget Office, 1981, 1-66.
- Poterba, James M.**, "Explaining the Yield Spread Between Taxable and Tax-Exempt Bonds: The Role of Expected Tax Policy," in Harvey S. Rosen, ed., *Studies in State and Local Public Finance*, Chicago: University of Chicago Press, 1986, 5-49.
- Rao, D. C. and Kaminow, Ira**, "Selective Credit Controls and the Real Investment Mix: A General Equilibrium Approach," in Ira Kaminow and James O'Brien, eds., *Studies in Selective Credit Policies*, Philadelphia, PA: Federal Reserve Bank of Philadelphia, 1975, 172-95.
- Riley, John G.**, "Credit Rationing: A Further Remark," *American Economic Review*, March 1987, 77, 224-7.
- Smith, Bruce**, "Limited Information, Credit Rationing, and Optimal Government Lending Policy," *American Economic Review*, June 1983, 73, 305-18.
- ____ and **Stutzer, Michael**, "Credit Rationing and Government Loan Programs: A Welfare Analysis," *AREUEA Journal* (American Real Estate and Urban Economics Association), Summer 1989, 17, 177-93.
- Stanton, Thomas H.**, *Government-Sponsored Enterprises: Their Benefits and Costs as Instruments of Public Policy*, Washington, DC: Association of Reserve City Bankers, 1988.
- Stiglitz, Joseph E. and Weiss, Andrew**, "Credit Rationing in Markets With Imperfect Information," *American Economic Review*, June 1981, 71, 393-410.
- Williamson, Stephen D.**, "Costly Monitoring, Loan Contracts, and Equilibrium Credit Rationing," *Quarterly Journal of Economics*, February 1987, 102, 135-45.
- Congressional Budget Office**, *New Approaches to the Budgetary Treatment of Federal Credit Activities*, Washington, DC: U.S. Government Printing Office, 1984.
- Council of Economic Advisers**, *Economic Report of the President*, Washington, DC: U.S. Government Printing Office, 1986.
- Flow of Funds Accounts**, Fourth Quarter 1987, Washington, DC: Board of Governors of the Federal Reserve System.
- General Accounting Office**, *Farmer's Home Administration: An Overview of Farmer Program Debts, Delinquencies, and Loan Losses*, Washington, DC: U.S. Government Printing Office, 1986.
- Office of Management and Budget**, "Special Analysis F," in *Special Analyses: Budget of the United States Government*, selected years.

Optimal Public Good Provision with Limited Lump-Sum Taxation

By JOHN DOUGLAS WILSON*

It is often argued that the use of distortionary taxation lowers the optimal provision of public goods below its optimal level in a first-best economy, which contains no restrictions on lump-sum taxation. However, this issue is usually investigated using commodity-tax models that contain no lump-sum taxes. This paper examines a many-consumer economy in which the only tax instruments are commodity taxes and a poll tax (subsidies are negative taxes). The optimal level of public good provision in this economy typically exceeds the first-best level, at least for distributionally neutral public goods. (JEL 321)

Much of the policy debate in recent years about the desirable level of public good provision in the United States seems at least implicitly to involve a comparison between first- and second-best public good levels. First-best policies are of little interest by themselves, since they apply to an economy in which there are no restrictions on the government's ability to employ nondistortionary lump-sum taxes to raise and redistribute income. However, they do serve as a useful benchmark against which the effects of limitations on the availability of lump-sum taxes can be characterized. Second-best policies reflect these limitations. A possible consensus among researchers on this comparison is reflected by the following statement from a well-known undergraduate text: "Since it becomes more costly to obtain public goods when taxation imposes distortions, normally this will imply that the efficient level of public goods is smaller than it would have been with nondistortionary taxation" (Joseph Stiglitz, 1988 p. 140).

It is now well understood by researchers that this claim needs to be carefully qualified. In particular, complementarities between public goods and taxed private goods could make it desirable to raise public spending beyond the first-best level. Fur-

thermore, particular public projects may have desirable consequences for the income distribution in a second-best economy but not in a first-best economy. Yet, if one is thinking about a "general bias," and if one concentrates on efficiency issues by considering only distributionally neutral public goods, then the assertion that the second-best public good level lies below the first-best level would seem to be a safe bet.¹

The basic message of the present paper is that this view must be modified when the availability of administratively simple forms of lump-sum taxation is recognized. It is standard practice to employ single-consumer models to analyze the efficiency issues associated with public good provision. These models allow the economy to be analyzed as though it contains a single consumer, either because all individuals are identical or because the income distribution is optimal. Distortionary commodity taxation is used to raise revenue, but only because the use of lump-sum taxation has been ruled out.² Moving to a many-con-

¹Even without complementarities or distributional problems, the second-best level may exceed the first-best level, but the examples in the literature, including my recent generalization (Wilson, 1990), support the "presumption" that this case normally will not happen.

²An income tax is usually modeled in this static framework as the special case of a commodity-tax system under which labor alone is taxed, or equivalently, all nonlabor commodities are taxed at identical

*Department of Economics, Ballantine Hall, Indiana University, Bloomington, IN 47405. David Starrett and the referees provided valuable comments.

sumer model, in which income-distribution considerations arise, it becomes possible to incorporate administratively feasible forms of lump-sum taxes into the model without destroying the desirability of using commodity taxation. One such lump-sum tax is a uniform poll tax, which is distributed equally across consumers and may be positive or negative.³ Indeed, a negative poll tax (i.e., a poll subsidy) is a component of the much studied and debated negative income tax, and it even surfaced during the 1972 U.S. Presidential Campaign under the name "demogrant."

A poll tax provides the government with a nondistortionary source of marginal finance. Even if the government chooses a negative poll tax, it is free to finance public goods at the margin by making this tax less negative. Thus, the use of a poll tax may eliminate any positive relation between commodity tax rates and public good levels, even though the distortionary commodity taxes are still present to take care of income-distribution problems. For this reason, the standard efficiency argument for setting the second-best public good level below the first-best level may no longer hold. In fact, an important special case of my model yields the opposite conclusion: the second-best level actually lies *above* the first-best level. This result does not rely on the income-distribution effects of public goods, complementarities between public and private goods, or unusual properties of the utility function, all of which are absent. In fact, my assumptions

encompass the case in which a Cobb-Douglas function determines the utility from private commodities, an assumption employed by Anthony Atkinson and Nicholas Stern (1974) and by Mervyn King (1986) to reach the opposite conclusion for a single-consumer model. The basic intuition behind the result is that greater public good provision lowers the total deadweight loss from taxation by transferring resources from the distorted private sector of the economy to the fully controlled public sector. This reduction in deadweight loss represents an additional benefit of public good provision that is absent in the first-best economy.

Moving to the general model introduces additional considerations into the comparison of first- and second-best public good levels. But as a description of a "general bias" for distributionally neutral public goods, the claim that second-best public good levels lie below first-best levels still appears to be unwarranted when the availability of administratively feasible forms of lump-sum taxation is recognized.

The inclusion of poll taxes in a model of optimal public good provision and commodity taxation is certainly not new. Peter Diamond (1975), Atkinson and Stiglitz (1980), Richard Tresch (1981), David Wildasin (1984), King (1986), and David Starrett (1988) all present rules for optimal public good provision in a many-consumer economy with commodity taxes and a poll tax. However, the issue of how distortionary taxation affects public good *levels* is distinct from the issue of how it affects the *rules* for optimal public good provision. Both issues are important, but the distinction between them is sometimes underappreciated. For the "levels issue," a global comparison of first- and second-best public good levels appears to have been limited to simple special cases of single-consumer models, no doubt reflecting recognized difficulties with obtaining general results in this area.

The plan of this paper is as follows. The next two sections formulate the government's maximization problem and derive the crucial optimality conditions. Section III then uses these conditions to derive the main proposition, which suggests that effi-

ad valorem rates. For intertemporal interpretations of the model, the commodity tax rates can be chosen to represent an income tax on both labor and interest income.

³Even a poll tax would entail some administrative costs, but the point here is that such costs would be relatively small. In his masterly survey of optimal tax theory, James Mirrlees (1986, p. 1197) asserts that "There is little difficulty about paying the same subsidy to every individual in the economy; there is not much more difficulty in making the subsidy depend on age. Uniform positive taxes may be a little more difficult. Taxes and subsidies proportional to trade in specified goods or service may also be difficult to administer with perfect accuracy. But, subject to some minor imperfections, we can take it that most such taxes use information that is cheaply and publicly available."

ciency considerations tend to raise the second-best public good level above the first-best level. The first example in Section IV supports this interpretation. In a second example, however, the second-best level lies below the first-best level, because the public good has important distributional effects. Yet the familiar efficiency argument for a relatively low second-best level does not appear in this second example, casting further doubt on its relevance. Section V draws some connections between the results reported here and those in related literature.

1. The Government's Optimization Problem

The economy contains H individuals, who differ in the form of their utility functions. Public goods are treated as a single aggregate quantity that is "pure" in the sense that all individuals receive the same amount, G , which can be provided to additional individuals at no cost (e.g., national defense). The assumption of a pure public good is employed in all of the previously cited references, but it does have important distributional implications, which are later discussed.

For the second-best economy under consideration, all individuals also face the same vector of commodity tax rates and pay the same poll tax, denoted T (possibly negative). Given this tax system and the public good supply, individual h chooses a vector of net demands for private commodities, $\mathbf{x} = (x_1, \dots, x_N)$, to solve the standard consumer problem of maximizing a strictly quasi-concave utility function, subject to a budget constraint:

$$\max_{\mathbf{x}} u^h(\mathbf{x}, G)$$

subject to

$$(1) \quad \sum_i q_i x_i \leq -T$$

where q_i is the consumer price of commodity i , equal to the producer price p_i plus the tax rate t_i . In vector notation, $\mathbf{q} = \mathbf{p} + \mathbf{t}$. The solution to h 's utility-maximization problem defines the demand functions, $x_i^h(\mathbf{q}, G, Y)$,

and the indirect utility function, $v^h(\mathbf{q}, G, Y)$, where Y is lump-sum income, equal to $-T$ in this model. One of the commodities, say 1, may be interpreted as "resources" (usually more specifically described as labor), in which case $x_1^h(\mathbf{q}, G, Y)$ is negative, because individuals are net suppliers of this commodity. Since only relative prices matter, both the consumer and producer price of commodity 1 may be set equal to 1.

This setup follows the common practice of assuming that there are no fixed factors in the economy. By eliminating this obvious nondistortionary source of finance for public expenditures, I place "severe" limits on the availability of lump-sum taxation. It can then be seen whether a very limited form of lump-sum taxation is enough to reverse the conclusions from the single-consumer analysis.

There is no need to assume that all commodities are taxed optimally. Rather, the government can be given control of a vector of "tax instruments," \mathbf{b} , which determines each tax rate according to a function $t_i(\mathbf{b})$. For example, these functions may constrain some commodities to be taxed at identical rates and others not to be taxed at all. They may even fix all tax rates at specific levels, $t_i(\mathbf{b}) = t_i^*$, with changes in the public good supply then being met with changes in the poll tax to balance the government budget. For our purposes, the critical feature of each of these special cases is that the poll tax and public good supplies can be varied independently of the commodity tax rates, at least in a neighborhood of the government's optimum.

The government's problem is to choose \mathbf{b} , T , and G to maximize a generalized utilitarian social welfare function, $\sum_h \Psi(u^h)$ for an increasing and concave function Ψ ,⁴ subject to the economy's production constraint. This constraint is assumed to be linear in private commodities, and public good production is characterized by a convex cost function, de-

⁴This commonly used form of the social welfare function encompasses the utilitarian and Rawlsian objectives as special cases. The assumption of additive separability is not needed for Proposition 1, but it simplifies one of the later examples.

noted $c(G)$. The linearity assumption fixes producer prices, thereby allowing consumer prices to be defined directly as functions of \mathbf{b} : $q_i(\mathbf{b}) = p_i + t_i(\mathbf{b})$.⁵ The government's problem then becomes

$$\max_{\mathbf{b}, T, G} \sum_h \Psi[v^h(\mathbf{q}(\mathbf{b}), G, -T)]$$

subject to

$$(2) \sum_{i,h} p_i x_i^h(\mathbf{q}(\mathbf{b}), G, -T) + c(G) \leq 0.$$

By Walras's law, production constraint (2) implies the government budget constraint.

Omitted from this setup is the requirement that the consumer budget set defined by (1) intersect each individual's "consumption set," which describes his possible net demand vectors (reflecting, for example, the minimum consumption levels required for subsistence). I basically assume that these constraints can be omitted because they are not binding. One simple violation of this assumption is the case in which all commodity tax rates are constrained to equal 0 and some consumers lack sufficiently large endowments of commodities (including time) to pay the poll tax needed to finance the desired level of public good provision. However, for reasonable assumptions about both the availability of commodity taxation and the degree of inequality aversion embodied in the social welfare function, all individuals can be expected to possess endowments that significantly exceed the optimal poll tax in value. The "central case" in Stern's (1976) extensive numerical calculations of optimal linear income-tax schedules includes a poll

subsidy (i.e., a negative poll tax) equal to 34 percent of the average before-tax income (see Stern, 1976 p. 152).

II. First-Order Conditions

To state the first-order conditions for the government's maximization problem, let λ denote the Lagrange multiplier on constraint (2) and use subscripts other than i to denote partial derivatives and a prime to denote a total derivative. The Lagrangian for the government's problem may then be differentiated with respect to G to obtain the first-order condition,

$$(3) \sum_h \beta^h MRS^h - \lambda \left[\sum_{i,h} p_i x_{iG}^h + c'(G) \right] = 0$$

where

$$(4) \beta^h = \Psi'(v^h) \cdot v_Y^h$$

and

$$(5) MRS^h = v_G^h / v_Y^h.$$

The term β^h is individual h 's "social marginal utility of consumption," and MRS^h is h 's marginal rate of substitution between the public good and lump-sum income. By differentiating the Lagrangian with respect to T , one obtains another important first-order condition:

$$(6) \sum_h \left[\beta^h - \lambda \sum_i p_i x_{iT}^h \right] = 0.$$

This section combines these two first-order conditions into an expression that can be used directly to compare first- and second-best public good levels. In accordance with the previous claim that the analysis applies even to the case of an arbitrarily fixed vector of commodity tax rates, no use is made of any first-order conditions for these tax rates.

As a first step, divide both sides of (5) by λ and add and subtract the sum

⁵If the linearity assumption is replaced with the weaker assumption of constant returns to scale and we also assume that all commodities are taxed optimally, then none of the first-order conditions presented here change. The assumption of optimal taxation effectively allows the government to set \mathbf{q} independently of \mathbf{p} , in which case the first-order conditions for optimal taxation and public good provision do not contain changes in producer prices (see Diamond and Mirrlees, 1971). However, the actual comparison of first- and second-best public good levels would now have to deal with differences in producer prices between the two equilibria.

$$\sum_{i,h} p_i x_{iY}^h \text{MRS}^h:$$

$$(7) \quad \sum_h \left[(\beta^h / \lambda) - \sum_i p_i x_{iY}^h \right] \text{MRS}^h - \sum_{i,h} p_i [x_{iG}^h - x_{iY}^h \text{MRS}^h] = c'(G).$$

The term in the second set of square brackets gives the marginal impact of G on h 's net demand for i , compensated by a reduction in income that keeps h on his initial indifference curve. In other words, this term is the derivative with respect to G of h 's compensated demand function, $s_i^h(\mathbf{q}, G, u)$:

$$(8) \quad s_{iG}^h = x_{iG}^h - x_{iY}^h \text{MRS}^h.$$

Thus, the second summation essentially involves only efficiency considerations, which depend only on substitution effects, not income effects.

In contrast, the first summation in (7) measures equity effects. Specifically, the first-order condition for T given by (6) implies that this summation is a covariance between the bracketed term and MRS^h . To interpret this covariance, first define

$$(9) \quad \gamma^h = \beta^h + \lambda \sum_i t_i x_{iY}^h.$$

Diamond (1975) calls γ^h the "social marginal utility of income," because it measures the amount by which social welfare rises when individual h receives another unit of income, taking into account the social value of h 's change in tax payments. Using the equality, $\sum_i (p_i + t_i) x_{iY}^h = 1$, which is implied by the consumer budget constraint, one finds that

$$(10) \quad (\beta^h / \lambda) - \sum_i p_i x_{iY}^h = (\gamma^h / \lambda) - 1.$$

Equations (10) and (6) imply that λ equals the average γ^h and that the first summation in (7) is actually a covariance between γ^h / λ and MRS^h . Then, (7) may be rewritten as a sum of this equity term and the efficiency

term obtained by using (8):

$$(11) \quad - \sum_{i,h} p_i s_{iG}^h(\mathbf{q}^{**}, G^{**}, u^{h**}) + \text{Cov}(\gamma^h / \lambda, \text{MRS}^{h**}) = c'(G^{**})$$

where double asterisks denote the second-best values of variables and the arguments of the compensated demand derivatives in (11) are explicitly included for later reference.

For guidance in the manipulations of (11) needed to compare public good levels, I now introduce a rule for the optimal public good supply in a first-best economy. Such a rule can be obtained directly from (11). With lump-sum taxes available to take care of distributional problems, the covariance drops out. The absence of commodity taxes implies that the first term should be evaluated at $\mathbf{q} = \mathbf{p}$, in which case this term becomes a sum over individuals of the amount by which each individual's expenditures on private commodities must rise to compensate for a unit drop in G . This compensation is the individual's marginal rate of substitution between G and Y :

$$(12) \quad \text{MRS}^h(\mathbf{p}, G, u^h) = - \sum_i p_i s_{iG}^h(\mathbf{p}, G, u^h).$$

Equation (11) therefore becomes the famous "Samuelson condition," requiring that the sum of the marginal rates of substitution equal the marginal cost of public good provision:

$$(13) \quad \sum_h \text{MRS}^h(\mathbf{p}, G^*, u^{h*}) = c'(G^*)$$

where single asterisks denote the first-best values of variables.

The next step in the comparison of public good levels is to decompose (11) into a sum of terms that includes the left side of (13) as one component. This task is accomplished by introducing deadweight-loss considerations explicitly into the analysis. Thus, de-

fine a "loss function" as follows:

$$(14) \quad L(\mathbf{q}, G, \mathbf{u}) = \sum_{i,h} p_i s_i^h(\mathbf{q}, G, u^h) \\ - \sum_{i,h} p_i s_i^h(\mathbf{p}, G, u^h)$$

where $\mathbf{u} = (u^1, \dots, u^H)$ and \mathbf{p} is omitted as an explicit argument of $L(\cdot)$ because it is fixed for the analysis.⁶ In words, this function gives the amount by which the resource cost of the net demands in the equilibrium with taxes $\mathbf{t} = \mathbf{q} - \mathbf{p}$ exceeds the minimum resource cost needed to achieve the given utility vector with the given G . Because taxes distort consumption decisions, this minimum resource cost is achieved only if \mathbf{q} is proportional to \mathbf{p} . Otherwise, the deadweight loss is positive (assuming some substitutability in consumption).

The marginal deadweight loss from an increase in G may now be defined as the partial derivative of the loss function with respect to G . In words, it is the marginal impact of G on the excess resource cost needed to achieve the current utility vector. By substituting this derivative of (14) into (11) and using (12), I obtain

$$(15) \quad \sum_h \text{MRS}^h(\mathbf{p}, G^{**}, u^{h**}) \\ - L_G(\mathbf{q}^{**}, G^{**}, \mathbf{u}^{**}) \\ + \text{Cov}(\gamma^h/\lambda, \text{MRS}^{h**}) = c'(G^{**}).$$

A comparison of (15) with (13) shows that a positive (negative) marginal impact of G on the deadweight loss from taxation contributes negatively (positively) to the marginal benefit of G . Complicating the comparison, however, are "income effects" created by the presence of u^{h**} in (15) and u^{h*} in (13). The impact of these utility differences on the marginal benefit of G may

be described by the difference

$$(16) \quad \sum_h \Delta \text{MRS}^h = \sum_h \text{MRS}^h(\mathbf{p}, G^{**}, u^{h**}) \\ - \sum_h \text{MRS}^h(\mathbf{p}, G^{**}, u^{h*}).$$

Substituting (16) into (15) gives the first-order condition that will be used for the comparison:

$$(17) \quad \sum_h \text{MRS}^h(\mathbf{p}, G^{**}, u^{h*}) \\ - L_G(\mathbf{q}^{**}, G^{**}, \mathbf{u}^{**}) \\ + \text{Cov}(\gamma^h/\lambda, \text{MRS}^{h**}) \\ + \sum_h \Delta \text{MRS}^h = c'(G^{**}).$$

III. The Main Comparison

The main proposition about first- and second-best public good levels will now be presented, proved, and discussed.

PROPOSITION 1: $G^{**} \geq G^*$ as

$$(18) \quad -L_G(\mathbf{q}^{**}, G^{**}, \mathbf{u}^{**}) \\ + \text{Cov}(\gamma^h/\lambda, \text{MRS}^{h**}) \\ + \sum_h \Delta \text{MRS}^h \geq 0.$$

PROOF:

If consumer prices are held fixed at \mathbf{p} (i.e., $\mathbf{t} = \mathbf{0}$), all individuals possess well-defined indifference curves over the public good and lump-sum income. The strict convexity of these indifference curves and the convexity of the cost function $c(G)$ then imply that

$$G^{**} \geq G^* \quad \text{as}$$

$$(19) \quad \left[\sum_h \text{MRS}^h(\mathbf{p}, G^{**}, u^{h*}) - c'(G^{**}) \right] \\ - \left[\sum_h \text{MRS}^h(\mathbf{p}, G^*, u^{h*}) - c'(G^*) \right] \leq 0.$$

⁶This definition is equivalent to J. A. Kay's (1980) definition of the loss function, which generalizes Diamond and Daniel McFadden's (1974) definition.

Subtracting (13) from (17) shows that (19) is equivalent to (18).

Thus, the difference $G^{**} - G^*$ depends positively on three effects: 1) a deadweight-loss effect, given by the negative of the derivative of the loss function with respect to G ; 2) an equity effect, given by the covariance between γ^h/λ and MRS^{h**} ; and 3) an income effect, given by the difference $\sum_h \Delta MRS^h$ in (16). For the purpose of uncovering a "general bias," it seems reasonable to consider the case in which the utility functions exhibit separability between private commodities and the public good, since "...a general project has no obvious net complementarity" (Starrett, 1988 p. 173). In this case, it can be argued intuitively that a marginal rise in G should normally lower the deadweight loss, thereby creating a tendency for G^{**} to exceed G^* . The separability assumption implies that the utility functions take the form, $\bar{u}^h(w^h(x), G)$, where $w^h(x)$ denotes "private utility." With commodity tax rates and total utilities held fixed, a rise in G affects the loss function [defined in (14)] only by reducing private utilities. In other words, a higher G may be viewed as representing a transfer of resources away from the distorted private sector and into the government-controlled public sector. With fewer resources employed in the sector that misallocates them, the deadweight loss should normally be lower.

In general, however, both the second and third effects are also present, and they may both contribute towards making G^{**} fall short of G^* . Assuming G is a strongly normal good, $\sum_h \Delta MRS^h$ is likely to be significantly negative, because the deadweight loss from the distortionary commodity taxes can be expected to lower utilities, at least on average. Under the same normality assumption, however, individuals with high utilities and low social marginal utilities of income are likely to place high marginal values on the public good, in which case the equity effect is also significantly negative. Example 2 in the next section illustrates this reasoning.

Conversely, if one considers a public good that is distributionally neutral, as defined by

a zero equity effect, then one should expect the income effect to be relatively small. In terms of a "general bias" then, Proposition 1 suggests that the government's use of distortionary taxation tends to *raise* the optimal level of distributionally neutral public goods above the first-best level. Example 1 in the next section supports this claim.

Some readers may be puzzled by the complete absence of terms reflecting increased commodity tax rates in Proposition 1, especially since such tax changes are central to the standard argument that the efficiency losses from taxation lower the optimal public good level below its first-best value. A standard envelope-theorem argument explains this absence: if the current G is being optimally financed with the available tax instruments, then the marginal source of finance is irrelevant in the sense that all methods of financing a marginal rise in G have the same welfare impacts. In particular, if poll-tax financing is available and the government is optimizing over available taxes, then it must be the case that the relatively large deadweight loss associated with commodity-tax finance is offset by desirable equity advantages at the margin, making commodity taxes and the poll tax equally attractive at the margin. One may then analyze the desirability of a marginal rise in G by arbitrarily assuming that poll taxes are used to finance it. Hence, the resulting condition contains no changes in commodity tax rates.

IV. Examples

Proposition 1 is now put to work on two interesting examples. In the first example, only the deadweight-loss effect from Proposition 1 is found to be nonzero, and it causes G^{**} to lie above G^* . The second example respecifies the utility function so that the public good is a strongly normal good. As a result, both the income and equity effects become negative, in accordance with the previous contention that the two effects are intimately related to each other. Moreover, these two effects outweigh the deadweight-loss effect, causing G^{**} to fall short of G^* . Taken as a whole, the

examples support the claim that arguments for reducing the second-best public good level below its first-best value must be based on equity considerations, not efficiency considerations.

Example 1: Individuals are endowed only with "resources," and the only innate differences among individuals are endowment differences. With commodity 1 representing resources, the direct utility function for individual h is assumed to take the form, $u[w(r^h + x_1, x_2, \dots, x_N) + f(G)]$, where r^h is his resource endowment, $r^h + x_1$ is then his gross demand for resources ($x_1 < 0$), and the function w is homogeneous of degree one. I will begin the analysis by describing the special forms of the uncompensated and compensated demand functions for this model. The uncompensated demands for consumer h are found by solving h 's utility-maximization problem:

$$(20) \quad \max_x u[w(r^h + x_1, x_2, \dots, x_N) + f(G)]$$

subject to

$$(r^h + x_1) + \sum_{i \geq 2} q_i x_i \leq r^h - T$$

where resources represent the untaxed numeraire (i.e., $q_1 = p_1 = 1$). Since the function w is homogeneous of degree one, the uncompensated demand functions possess the following forms for some function $\theta_i(q)$:

$$(21a) \quad x_i^h(q, G, -T) = (r^h - T)\theta_i(q) \quad i \geq 2$$

$$(21b) \quad x_1^h(q, G, -T) = (r^h - T)\theta_1(q) - r^h.$$

In other words, the net demands do not depend on G and are linear in "full income," defined as $r^h - T$.

To obtain the compensated demand functions, first substitute the uncompensated demand functions back into the direct utility function, giving the indirect utility function:

$$(22) \quad v^h(q, G, -T) = u[(r^h - T)\omega(q) + f(G)]$$

where $\omega(q) = w[\theta_1(q), \dots, \theta_H(q)]$ is the private utility from one unit of full income. Letting $\Omega(G, u)$ represent the private utility needed to achieve total utility u under public good supply G , one may then substitute for $r^h - T$ in (21) to obtain compensated demand functions with the following forms:

$$(23a) \quad s_i^h(q, G, u) = \frac{\Omega(G, u)}{\omega(q)} \theta_i(q) \quad i \geq 2$$

$$(23b) \quad s_1^h(q, G, u) = \frac{\Omega(G, u)}{\omega(q)} \theta_1(q) - r^h.$$

The previous section argued that the loss-function derivative in Proposition 1, $L_G(q, G, \underline{u})$, can be expected to be negative. To check this, substitute (23) into (14) and differentiate:

$$(24) \quad L_G(q, G, \underline{u}) = \sum_h \Omega_G(G, u^h) \times \left\{ \sum_i p_i [\theta_i(q)/\omega(q) - \theta_i(p)/\omega(p)] \right\} = \frac{\sum_h \Omega_G(G, u^h)}{\sum_h \Omega(G, u^h)} L(q, G, \underline{u}).$$

This expression is indeed negative, since commodity taxes create a positive deadweight loss, and private utility $\Omega(G, u^h)$ must fall to compensate for a rise in G .

Thus, G^{**} will exceed G^* , if the income and equity effects described in Proposition 1 do not offset the negative deadweight-loss effect. By (22), the marginal rate of substitution between full income and the public good is $f'(G)/\omega(q)$, which is independent of the value of full income. Since full income is the only dimension along which individuals differ, it follows that the equity effect described by the covariance in (18) equals 0, as does the income effect.

Thus, deadweight-loss considerations prevail, implying that the second-best public good level exceeds the first-best level.

Example 2: To introduce equity effects into the analysis, respecify the first example so that the private utility function continues to be homogeneous of degree one but preferences between private utility and the public good are Cobb-Douglas:

$$(25) \quad u^h(\mathbf{x}, G) \\ = \alpha \log[w(r^h + x_1, x_2, \dots, x_N)] \\ + (1 - \alpha) \log G$$

where $0 < \alpha < 1$. It is worth noting that this specification may be reinterpreted as one in which individuals differ in their abilities to perform labor, if one also assumes a Cobb-Douglas function for private utility. In particular, assume two commodities and write

$$(26) \quad u^h(x_1, x_2, G) \\ = \alpha \log[(r^h + x_1)^\mu (x_2)^{1-\mu}] \\ + (1 - \alpha) \log G$$

where $0 < \mu < 1$. Using the properties of logs, (26) becomes

$$(27) \quad u^h(x_1, x_2, G) \\ = \alpha \log[(1 + x_1/r^h)^\mu (x_2)^{1-\mu}] \\ + \alpha \mu \log r^h + (1 - \alpha) \log G.$$

Since the term $\alpha \mu \log r^h$ is a constant, it can be dropped without altering the analysis in any way, resulting in a model in which x_2 represents consumption and the h th individual works $-x_1/r^h$ hours to earn before-tax income $-x_1$. This specification of individual differences is standard in the optimal-income-tax literature, as originally developed by James Mirrlees (1971). Furthermore, the assumption of Cobb-Douglas preferences for private utilities underlies the single-consumer examples of Atkinson and Stern (1974) and King (1986).

Returning to the more general case of (25), note that (21) and (23) continue to describe the uncompensated and compen-

sated demand functions. Thus, (24) continues to describe the change in deadweight loss. However, this change must now be compared with nonzero income and equity effects. To investigate their signs, note first that the indirect utility function must be modified to account for Cobb-Douglas preferences between private utility and the public good. In particular, substituting (21) into (25) gives

$$(28) \quad v^h(\mathbf{q}, G, -T) \\ = \alpha \log[(r^h - T)\omega(\mathbf{q})] + (1 - \alpha) \log G$$

where $\omega(\mathbf{q})$ again denotes the private utility from one unit of full income. The marginal rate of substitution between the public good and full income is then

$$(29) \quad MRS^h = \frac{1 - \alpha}{\alpha G} (r^h - T).$$

Thus, MRS^h increases with full income. Since the social marginal utility of income declines with full income,⁷ it follows that the covariance term describing the equity effect is negative. A more complex argument, provided in the Appendix, uses (29) to show that the income effect is also negative. In fact, it alone outweighs the deadweight-loss effect, making $G^{**} < G^*$.

Thus, abandoning the assumption of distributional neutrality allows one to construct a reasonable example in which the second-best public good level falls short of the first-best level. Unlike the traditional efficiency argument for contracting the public good supply, this example is based on equity considerations. Moreover, it is important to recognize the limitations of the model which underlie the result, particularly the assumption of a pure public good. If the defining characteristics of public goods

⁷To see this, recall that the social marginal utility of income is $\gamma^h = \beta^h + \lambda \sum_i t_i x_{iy}^h$. The term β^h declines with full income under the Cobb-Douglas assumption, and (21) implies that $\sum_i t_i x_{iy}^h = \sum_i t_i \theta_i(\mathbf{q})$, which is independent of full income. Thus, γ^h declines with full income.

were generalized to allow for limited differences in public good levels across individuals, then such differences could substitute for a system of optimal lump-sum transfers, albeit imperfectly. A more expanded definition of public goods may therefore create an equity argument for greater public good provision in a second-best economy than in a first-best economy. For future research, it would be useful to treat separately different broad categories of public goods when comparing the first- and second-best optima.

V. Related Literature

I have shown that the existence of nondistorting sources of government revenue necessitates a rethinking of how distortionary taxes affect the optimal supply of public goods. David Wildasin (1987a, b) makes a similar point when he criticizes the common belief that the ability of local governments to "export" taxes leads to an overprovision of public goods from the viewpoint of the nation's welfare. This belief runs into trouble when there exist some tax instruments that do not involve exporting (e.g., taxes on nontraded commodities demanded or supplied only by residents). When a local government optimizes over all of its tax instruments, it is indifferent at the margin between "exported" and "own-source" revenues. The possibility of tax exporting then becomes irrelevant for marginal public-expenditure decisions. As a result, Wildasin is able to present an example in which the elimination of tax exporting has no impact on the optimal public good level. The poll tax in the present paper acts in a manner similar to Wildasin's "own-source" revenues by eliminating the negative impact of commodity taxation on the optimal level of public goods.

A real issue in both Wildasin's work and the present paper is whether it is legitimate to assume that governments pursue welfare-maximizing tax policies, given the available tax instruments. If they do not, then there could be a sizable welfare difference between various types of tax finance at the margin. Unfortunately, as Edgar Browning (1976 p. 296) points out, "It is

important to recognize that it is literally impossible to determine the exact source of funds when governments use general-fund financing (enacting tax and expenditure bills separately)." Such a determination is critical for benefit-cost analysis, for which the tax side of government is traditionally taken as exogenous to the problem and the researcher is concerned only with formulating rules for the evaluation of particular public projects. However, the intent of the present paper has not been to treat the tax side as an accurate description of existing U.S. tax policies, but rather to include the tax side as part of the social-welfare optimization problem confronting the government. Such an approach seems appropriate for evaluating the common belief that problems with tax administration depress the desirable level of public good provision by limiting the use of nondistortionary taxes.

Several articles listed at the beginning of this paper discuss how the first- and second-best benefit-cost rules differ in a many-consumer economy. Unfortunately, this literature does not help much with determining how first- and second-best public good levels differ. Atkinson and Stern (1974) thoroughly depict the distinction between the "levels issue" and the "rules issue" in a single-consumer economy with commodity-tax financing. To see the importance of this distinction within the context of the current paper's many-consumer model, rewrite equation (11) as follows:

$$\begin{aligned}
 (30) \quad & \sum_h \text{MRS}^h(\mathbf{q}^{**}, G^{**}, u^{h**}) \\
 & + \sum_{i,h} t_i s_{iG}^h(\mathbf{q}^{**}, G^{**}, u^{h**}) \\
 & + \text{Cov}(\gamma^h/\lambda, \text{MRS}^{h**}) = c'(G^{**})
 \end{aligned}$$

where use is made of the equality

$$\begin{aligned}
 (31) \quad & \text{MRS}^h(\mathbf{q}^{**}, G^{**}, u^{h**}) \\
 & = - \sum_i q_i s_{iG}^h(\mathbf{q}^{**}, G^{**}, u^{h**})
 \end{aligned}$$

(i.e., h 's marginal rate of substitution equals the amount by which his expenditures must rise to maintain his initial utility level following a marginal fall in G). If commodity 1 ("resources") again serves as the numeraire and all taxes are positive, then the second term in (30) is necessarily negative in the first example [see eq. (23a)]. If one ignores the equity term, which is 0 in this same example, then it follows that

$$(32) \quad \sum_h MRS^h(q^{**}, G^{**}, u^{h**}) > c'(G^{**}).$$

In other words, public good expenditures should be carried out to the point at which their marginal benefit, as conventionally measured, still exceeds the marginal cost. King (1986) argues that this result holds in many cases; but it cannot imply that G^{**} lies below G^* , because we have derived the opposite result for the same model.⁸

This apparent conflict in results is resolved by observing that the difference between the optimal public good levels depends on the marginal impact of G on the deadweight loss from taxation, whereas the difference between the two sides of (32) depends on the second term in (30), which has no relation to the deadweight loss measure. As King (1986 p. 283) succinctly describes it, this second term "measures the distortion in the aggregate willingness to pay [for G] resulting from the use of distorting taxes to finance government expenditures." Under the assumptions of the first example, this "distortion" reflects the loss in commodity tax revenue that the government experiences when resources are transferred from the private sector to the public sector via the poll tax. Since individuals do not take this lost revenue into account, they

overvalue the public good from the viewpoint of social welfare.⁹

A final literature that deserves some discussion consists of the many papers that numerically compute the "marginal welfare cost of taxation." Recent examples include Charles Ballard et al. (1985), Browning (1987), and Charles Stuart (1984), all of which have recently been reviewed by Don Fullerton (1989). This literature ignores equity issues, concentrating instead on the marginal deadweight loss (or "marginal excess burden") from various distortionary methods of financing public expenditures at the margin. It generally obtains sizable marginal-loss measures, but Fullerton observes that only Browning's paper, among the three, provides a traditional measure of this marginal loss. He also argues, however, that this measure has limited relevance for benefit-cost analysis, because it does not determine the extent to which the conventional marginal-benefit measure should exceed the conventional marginal-cost measure in second-best economies, a view which is consistent with equation (30) above. Whether it is relevant for the comparison of first- and second-best public good levels depends crucially on the availability of a poll tax.

To understand this last distinction, alter first-order condition (17), on which Proposition 1 is based, to make it applicable to a single-consumer economy. The existence of a single representative consumer eliminates the covariance term, the summation signs, and the superscripts identifying individuals; and the requirement that all government revenue needs be met by distortionary com-

⁸King (1986) later presents a Cobb-Douglas example in which the second-best public good level falls short of the first-best level. However, for this example, he switches to a single-consumer model without poll taxes. As has been seen, it is dangerous to extrapolate the results from single-consumer models to models with many consumers.

⁹Unfortunately, King (1986) also refers to the second term in (30) as the "Pigou term" and claims that it measures the "indirect damage" caused by taxation. A. C. Pigou (1949 pp. 33-4) was concerned with how benefit-cost analysis should account for the distortions created by the taxes needed to finance additional public expenditures. The modern public-economics literature uses loss-function derivatives to measure the cost of these distortions. As explained at the end of Section II, no such derivatives appear in my rules for the optimal public good supply, because a nondistortionary poll tax is available to finance the public good at the margin.

modity taxation is taken into account by supplementing the deadweight-loss term in (17) with an expression measuring the marginal deadweight loss from commodity taxation. With these changes, the optimal G for a single-consumer economy must satisfy

$$(33) \quad \text{MRS}(\mathbf{p}, G^{**}, u^*) - L_G(\mathbf{q}^{**}, G^{**}, u^{**}) \\ - L_q(\mathbf{q}^{**}, G^{**}, u^{**})(dt/dG) \\ + \Delta \text{MRS} = c'(G^{**})$$

where dt/dG denotes the vector of tax changes used to finance G at the margin (as derived from the government's optimal-tax problem).¹⁰ Differentiation of equation (14) shows that the k th element of the vector of partial derivatives, L_q , is $-\sum_i t_i (\partial s_i / \partial q_k)$. Browning confines his analysis to a tax only on labor, which may be represented here by a negative value of $t_1 = q_1 - p_1$, reflecting a subsidy on leisure. In this case, implicit differentiation of the government budget constraint gives the tax change that finances a dollar increase in public good expenditures from the optimum, and one can write the marginal deadweight loss from this tax change as follows:

$$(34) \quad \frac{\partial L(\mathbf{q}^{**}, G^{**}, u^{**})}{\partial q_1} \left(\frac{dt_1}{dG} \right) [c'(G^{**})]^{-1} \\ = \frac{\tau \eta_1}{1 - \tau \eta_1} (1 - \tau \eta_G \varphi)$$

where

$$\tau = -\frac{t_1}{q_1} \quad \eta_1 = \left(\frac{\partial s_1}{\partial q_1} \right) \left(\frac{q_1}{s_1} \right) \\ \eta_G = \left(\frac{\partial s_1}{\partial G} \right) \left(\frac{G}{s_1} \right) \quad \varphi = \frac{-q_1 s_1}{c'(G) \cdot G}$$

with all terms evaluated at their second-best values.¹¹ Browning's numerical calculations

consistently produce sizable values for this type of marginal loss.¹² Because of its presence in the single-consumer economy, it is possible to show that $G^{**} < G^*$ for the types of utility functions considered in Section IV, given the added assumption of a CES (constant elasticity of substitution) private utility function and some mild restrictions on parameter values (see Wilson, 1990).

However, the first example in Section IV gives $G^{**} > G^*$ for the many-consumer case, in which a poll tax is allowed. Moreover, the assumptions in this example do not limit the size of the marginal loss in (34), since they do not restrict the wage elasticity η_1 . Therefore, it appears that any general tendency for the marginal deadweight losses from higher commodity taxes to depress the optimal public good supply vanishes once a reasonable form of lump-sum taxation is introduced into the economy.

APPENDIX

This appendix shows that the income effect in Example 2 is negative and exceeds the deadweight-loss effect in magnitude. To make this comparison, first transform the deadweight-loss effect given by (24) into a more useful expression by implicitly differentiating (28) to solve for the marginal rate of substitution between the public good and private utility:

$$-\Omega_G(G, u^h) = [(1-\alpha)/(\alpha G)] [\Omega(G, u^h)].$$

ancing tax and expenditure change that originates from the optimum has a zero first-order impact on utility. The elasticity η_1 is necessarily positive, since $s_1 < 0$ and $\partial s_1 / \partial q_1 < 0$.

¹²Browning's "preferred estimates" range from 0.318 to 0.469, but both numbers reflect the assumption that the marginal tax rate on labor increases faster than the average tax rate. The only other real difference between (34) and Browning's equation 11 is the presence of the elasticity η_G , which equals 0 under his assumption that the public good and private income are perfect substitutes. Browning also effectively considers the case in which the uncompensated demand derivatives, $\partial x_1 / \partial G$ and $\partial x_1 / \partial q_1$, both equal 0. Here, the government budget constraint gives $dt_1 / dG = c'(G) / x_1$, which reduces the right side of (34) to $\tau \eta_1$, an expression equivalent to Browning's equation 10.

¹⁰Formal proofs of equations (33) and (34) are available from the author upon request.

¹¹Equation (34) can be written in terms of compensated labor-supply elasticities, because any budget-bal-

Substituting this equality into (24) gives

$$(A1) \quad L_G(\mathbf{q}^{**}, G^{**}, \underline{\mathbf{u}}^{**}) \\ = -[(1-\alpha)/(\alpha G)] L(\mathbf{q}^{**}, G^{**}, \underline{\mathbf{u}}^{**})$$

where the double asterisk again denotes the second-best values of variables.

Equation (29) shows that the income effect, as defined by (16), is created by the change in each individual's full income needed to move his utility from its first-best level to its second-best level, given the second-best public good level and no commodity taxes. Letting ΔF^h denote this income change for individual h ,

$$(A2) \quad \sum_h \Delta \text{MRS}^h \\ = [(1-\alpha)/(\alpha G)] \sum_h \Delta F^h.$$

Thus, $\sum_h \Delta F^h$ in (A2) must be compared with $L(\mathbf{q}^{**}, G^{**}, \underline{\mathbf{u}}^{**})$ in (A1).

With consumer prices equaling producer prices in the absence of commodity taxes, the consumer budget constraints require that

$$(A3) \quad \sum_h \Delta F^h = \sum_{i,h} p_i [s_i^h(\mathbf{p}, G^{**}, u^{h**}) \\ - s_i^h(\mathbf{p}, G^{**}, u^{h*})].$$

Then,

$$(A4) \quad \sum_h \Delta F^h = \sum_{i,h} p_i [s_i^h(\mathbf{p}, G^{**}, u^{h**}) \\ - s_i^h(\mathbf{q}^{**}, G^{**}, u^{h**})] \\ + \sum_{i,h} p_i [s_i^h(\mathbf{q}^{**}, G^{**}, u^{h**}) \\ - s_i^h(\mathbf{p}, G^{**}, u^{h*})].$$

Using the definition of the loss function given by (14) in the text, (A4) may be

rewritten as

$$(A5) \quad \sum_h \Delta F^h = -L(\mathbf{q}^{**}, G^{**}, \underline{\mathbf{u}}^{**}) \\ + \sum_{i,h} p_i [s_i^h(\mathbf{q}^{**}, G^{**}, u^{h**}) \\ - s_i^h(\mathbf{p}, G^{**}, u^{h*})].$$

To sign the summation on the right side of (A5), observe first that the second-best optimum must satisfy the economy's production constraint [eq. (2)] with equality:

$$(A6) \quad \sum_{i,h} p_i s_i^h(\mathbf{q}^{**}, G^{**}, u^{h**}) = -c(G^{**}).$$

Furthermore, the nonoptimality of G^{**} in the first-best economy implies that

$$(A7) \quad \sum_{i,h} p_i s_i^h(\mathbf{p}, G^{**}, u^{h*}) \geq -c(G^{**}).$$

(Otherwise, there would exist a feasible way to change G and transfer income across consumers so as to make everyone better off.) Thus, the summation in (A5) is negative, implying that

$$(A8) \quad \sum_h \Delta F^h \leq -L(\mathbf{q}^{**}, G^{**}, \underline{\mathbf{u}}^{**}).$$

Using (A8) to compare (A2) with (A1) gives

$$(A9) \quad \sum_h \Delta \text{MRS}^h < L_G(\mathbf{q}^{**}, G^{**}, \underline{\mathbf{u}}^{**}) < 0$$

(i.e., the income effect is negative and outweighs the deadweight-loss effect). Since the text notes that the equity effect is also negative, Proposition 1 shows that G^{**} must fall short of G^* .

REFERENCES

- Atkinson, Anthony B. and Stern, Nicholas H., "Pigou, Taxation, and Public Goods," *Review of Economic Studies*, January 1974, 41, 119-28.
 _____ and Stiglitz, Joseph E., *Lectures on*

- Public Economics*, New York: McGraw-Hill, 1980.
- Ballard, Charles L., Shoven, John B. and Whalley, John, "General Equilibrium Computations of the Marginal Welfare Costs of Taxes in the United States," *American Economic Review*, March 1986, 75, 128-38.
- Browning, Edgar K., "The Marginal Cost of Public Funds," *Journal of Political Economy*, April 1976, 84, 283-98.
- _____, "On the Marginal Welfare Cost of Taxation," *American Economic Review*, March 1987, 77, 11-23.
- Diamond, Peter A., "A Many-Person Ramsey Tax Rule," *Journal of Public Economics*, November 1975, 4, 335-42.
- _____, and McFadden, Daniel L., "Some Uses of the Expenditure Function in Public Finance," *Journal of Public Economics*, February 1974, 3, 3-21.
- _____, and Mirrlees, James A., (1971a) "Optimal Taxation and Public Production I: Production Efficiency," *American Economic Review*, March 1971, 61, 8-27.
- _____, and _____, (1971b) "Optimal Taxation and Public Production II: Tax Rules," *American Economic Review*, June 1971, 61, 261-78.
- Fullerton, Don, "If Labor Is Inelastic, Are Taxes Till Distorting?" NBER Working Paper No. 2810, January 1989.
- Kay, J. A., "The Deadweight Loss from a Tax System," *Journal of Public Economics*, February 1980, 13, 111-9.
- King, Mervyn A., "A Pigovian Rule for the Optimum Provision of Public Goods," *Journal of Public Economics*, August 1986, 30, 273-91.
- Mirrlees, James A., "An Exploration in the Theory of Optimum Income Taxation," *Review of Economic Studies*, April 1971, 38, 175-208.
- _____, "The Theory of Optimal Taxation," in K. J. Arrow and M. D. Intriligator, eds., *Handbook of Mathematical Economics*, Vol. III, Amsterdam: North-Holland, 1986, 1197-1249.
- Pigou, A. C., *A Study in Public Finance*, 3rd (revised) Ed., London: Macmillan, 1949.
- Starrett, David A., *Foundations of Public Economics*, New York: Cambridge University Press, 1988.
- Stern, Nicholas H., "On the Specification of Models of Optimum Income Taxation," *Journal of Public Economics*, July-August 1976, 6, 123-62.
- Stiglitz, Joseph E., *Economics of the Public Sector*, New York: Norton, 1988.
- Stuart, Charles, "Welfare Costs per Dollar of Additional Tax Revenue in the United States," *American Economic Review*, June 1984, 74, 352-62.
- Tresch, Richard W., *Public Finance: A Normative Theory*, Plano, TX: Business Publications, 1981.
- Wildasin, David E., "On Public Good Provision with Distortionary Taxation," *Economic Inquiry*, April 1984, 22, 227-43.
- _____, (1987a) "Tax Exporting and the Marginal Cost of Public Expenditure," *Economics Letters*, 1987, 24 (4), 353-8.
- _____, (1987b) "The Demand for Public Goods in the Presence of Tax Exporting," *National Tax Journal*, December 1987, 40, 591-601.
- Wilson, John D., "Optimal Public Good Provision in the Ramsey Tax Model: A Generalization," *Economics Letters*, 1990 (forthcoming).

Retrospective Capital Gains Taxation

By ALAN J. AUERBACH*

This paper presents a new approach to the taxation of capital gains that eliminates the deferral advantage of realization-based systems, along with the lock-in effect and tax-arbitrage possibilities associated with this deferral advantage. The new method still taxes capital gains only upon realization but, effectively by charging interest on past gains when realization finally occurs, eliminates the incentive to defer such realization. Unlike a similar scheme suggested previously by Vickrey, the present method does not require knowledge of the potentially unobservable pattern of gains over time. It thus is applicable to a very broad range of capital assets. (JEL 320, 520)

Virtually every country that taxes income imposes a capital gains tax only upon the realization of gains rather than on accrual. Though countries vary with respect to indexing for inflation and the relative tax rates on capital gains and ordinary income, the realization-based tax system sets capital gains taxation apart from other forms of taxation and is associated with a variety of economic distortions.

The most frequently discussed problem arising from taxing capital gains upon realization is the "lock-in" effect, the desire to hold appreciated assets in order to defer taxes on gains already accrued. This effect leads investors to accept a lower before-tax rate of return than they would for new investments without such accrued gains, resulting in a distorted allocation of capital and inefficient portfolio selection.

As an illustration of the lock-in effect, consider a simple two-period example with-

out uncertainty in which an investor, having accrued a first-period gain, g , must decide whether to realize the gain and reinvest at the rate of return, i , or hold the asset for an additional rate of return r . Assuming all capital income is taxed at the same rate, t , then the investor's terminal wealth under the first strategy is

$$\begin{aligned}(1) \quad W_R &= [1 + g(1 - t)][1 + i(1 - t)] \\ &= (1 + g)(1 + i) \\ &\quad - t\{g[1 + i(1 - t)] + (1 + g)i\}.\end{aligned}$$

In second-period units, total taxes equal those paid in the first period, accumulated at the net-of-tax interest rate, plus those due in the second period.

If the investor chooses to hold rather than sell, the terminal wealth is

$$\begin{aligned}(2) \quad W_H &= (1 + g)(1 + r) \\ &\quad - t[(1 + g)(1 + r) - 1] \\ &= (1 + g)(1 + r) - t[g + (1 + g)r]\end{aligned}$$

so that the tax on the first-period gain is deferred, *without interest*, to the second period. This makes the investor willing to hold even for a range of returns $r < i$. The larger is g , the larger the deferral advantage and, hence, the lower r must be to induce the investor to sell.

*Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104-6297. Research support of the National Science Foundation and the Penn Institute for Law and Economics is gratefully acknowledged. Work on the paper began during a visit to the Financial Markets Group at the London School of Economics. I am grateful to Albert Ando, Doug Bernheim, Jim Hines, Bob McDonald, Jim Poterba, Jeff Strnad, Al Warren, seminar participants at the Universitat Autònoma de Barcelona, Columbia, NBER, Northwestern, Penn, Princeton, and Queens and three anonymous referees for comments on earlier drafts.

Closely related to the lock-in effect is the general problem of tax avoidance facilitated by the voluntary nature of realization. Because losses as well as gains have their tax burdens deferred until realization, investors have the incentive to realize losses immediately, to maximize the associated tax reductions. Aggressive application of the simple rule of holding winners and realizing losers potentially permits individuals to generate tax reductions without incurring major transaction costs (George M. Constantinides, 1983; Joseph E. Stiglitz, 1983).

This arbitrage possibility has led to a second major distortion arising from the present system of capital gains taxation. To prevent investors from generating capital losses to offset ordinary income, tax systems typically limit the allowable annual deduction for such losses. While perhaps representing an effective response to the problem of tax arbitrage, this loss-offset limitation also distorts the choice of investment away from the risky assets which are more likely to produce losses (e.g., Stiglitz, 1969).

Given such problems, there is great appeal to the prospect of switching to a tax on accrued capital gains. Though proposals to adopt accrual taxation have received serious scholarly attention (e.g., David Shakow, 1986), there seems to be little chance that such a system will be adopted on a broad scale.¹ Beyond the criticism that accrual taxation would increase annual taxpayer compliance costs, perhaps the most significant arguments against it are that some assets are hard to value except when they are sold and that liquidity constraints could force the premature sale of indivisible assets simply to pay the accruing taxes.

A potential solution to the problems of both present law and accrual taxation is a realization-based tax that offsets the deferral advantage of holding gains by imposing a higher tax rate on gains held for longer periods of time. The effect is to simulate a

system under which capital gains taxes are computed on an accrual basis but collected, with interest, only upon realization. From a comparison of (1) and (2), it is clear that charging tax-deductible interest on the taxes accruing on unrealized gains would eliminate the deferral advantage. Such an approach was originally conceived by William Vickrey (1939). By construction, it would eliminate the lock-in effect and the tax-arbitrage possibilities generated by selective realization, because of its equivalence to an accrual tax. At the same time, it would also remedy the liquidity problem of accrual taxation by collecting the tax only when sales actually occurred.

Unfortunately, this "cumulative averaging" approach is plagued by the second problem of accrual taxation mentioned above, that of valuation. For assets that are hard for the government to value except when sold, it will be unclear upon sale what the time pattern of accrual of the realized gain was. This will make it impossible to compute retrospectively the tax liability equal in present value to an annual tax on the asset's accrued gains (Jerry R. Green and Eytan Sheshinski, 1978). For example, if an asset has increased in value over a ten-year period, the tax rate on the realized gain needed to simulate accrual taxation would be the ordinary tax rate if the gain occurred entirely in the tenth year, but this tax rate compounded by one plus the relevant interest rate to the ninth power if the entire gain occurred during the first year of ownership. Simply to assume, for tax purposes, that a realized gain accrued smoothly at a constant annual rate would not solve the problem. Assets achieving above-normal rates of return initially would still be subject to a lock-in effect, because an investor anticipating only normal returns from the asset in the future would be able to spread the accrual pattern retrospectively imputed for this gain over several years by holding on to the asset. Likewise, an asset that had declined in value would offer its owner the incentive to sell. Thus, basic arbitrage transactions involving the holding of winners and the sale of losers would still be attractive, though perhaps less so than under a pure realization-based tax.

¹The tax system already has elements that effect accrual taxation, such as the mark-to-market requirements instituted in the 1981 Economic Recovery Tax Act to reduce tax-arbitrage activity involving commodity straddles.

Clearly, many capital assets, such as common shares of large companies, could be marked to market each year to avoid the valuation problem. However, an effective method of dealing with hard-to-value assets would still be necessary to make a switch to accrual taxation or accrual-equivalent realization taxation practical. This paper presents such a method.

The new approach does not require any information on the past pattern of accrued gains and yet eliminates the lock-in effect and the benefits of deferral-based tax arbitrage. In place of the private information on the accrued gains of individual assets, the scheme uses public information, the market interest rate, combined with the assumption of optimal portfolio choice by investors. It does not impose a uniform effective tax rate on accrued gains, *ex post*, but it does impose the same tax rate, *ex ante*, adjusted for risk.

The next section formalizes the criterion that a capital gains tax must satisfy in order not to distort the holding-period decision or allow deferral-based arbitrage. To provide the basic intuition about the new scheme and how it works, Section II presents the results for a special class of assets (such as precious metals) that generate no cash flows or tax liabilities until they are sold. Section III presents the solution for the general class of assets, while Section IV discusses certain limitations of the approach. Section V offers some concluding remarks.

I. Holding-Period Neutrality

The present system of taxation upon realization distorts behavior because the rate at which it taxes the income arising from an asset depends on the size of the asset's previous unrealized gains. This induces both the lock-in effect and deferral-related tax arbitrage.

Suppose that the risk-free interest rate is i ,² and the investor's tax rate on all forms of

income, including realized capital gains, is t . As shown in the example above, an investor holding an appreciated asset will require a before-tax return less than i from that asset to achieve his after-tax opportunity cost of $i(1-t)$, because the tax rate t applied to new gains is offset by the continued deferral, without interest, of taxes payable on the gains already generated but not yet realized.

This distortion would not be present under an accrual tax, which would tax additional income at the same rate regardless of unrealized appreciation or holding period. The result would be a required rate of return independent of these other characteristics. It is this result that we refer to as "holding-period neutrality."

Definition: A realization-based tax system is *holding-period neutral* if it leads each investor in an asset to require a before-tax return having a certainty-equivalent value that is not a function of the length of holding period or the asset's past pattern of returns.

As will be shown below, Vickrey's system of cumulative averaging satisfies this criterion. The challenge is to identify another system with weaker informational requirements that does so as well.

II. Retrospective Taxation

Beginning with a simple case will be helpful. The asset treated in this section is one that generates no cash flows or tax liabilities until it is sold and is taxed only upon sale. Henceforth, the analysis will be in continuous time.

At each date s , the investor (though not necessarily the government) is assumed to know the value of the asset, but not the asset's instantaneous return. Let $V(\cdot)$ be a valuation operator at each date that converts that date's distribution of uncertain returns into their certainty equivalents, from the investor's perspective.³

²If the tax system is not indexed for inflation, then this rate should be viewed as a nominal interest rate. Moreover, in the absence of a risk-free asset, one may reinterpret the paper's results in terms of a "zero-beta" asset that carries no risk premium.

³One can derive the function $V(\cdot)$ from a number of models for the valuation of risky returns, but the model

Consider the class of tax schemes that impose a tax T_s on realization (with $T_0 = 0$). A Vickrey-type tax system would satisfy

$$(3) \quad \dot{T}_s = i(1-t)T_s + tg_s A_s$$

where t is the rate of tax, i is the risk-free interest rate, g_s is the actual, *ex post*, rate of return on the asset at time (after purchase) s , and A_s is the asset's value at date s . As already indicated, though, the tax system described in (3) cannot be imposed retrospectively without knowledge of the time pattern of gains g_s . However, this expression is not a *necessary* condition for a holding-period-neutral tax. The fact that individual decisions are influenced by *ex ante* distributions of returns rather than by *ex post* returns allows us to pursue a weaker condition.

As the scheme in (3) is holding-period neutral (this will be shown formally below) and imposes the rate of tax t on newly accrued capital gains, the investor will require the same before-tax return on the capital asset and the safe asset, adjusting for risk: the certainty-equivalent value of the capital gain g_s will simply equal the investor's before-tax opportunity cost, $V(g_s) = i$. Applying $V(\cdot)$ to both sides of (3) yields

$$(4) \quad V(\dot{T}_s) = i(1-t)T_s + tiA_s.$$

Expression (4) says that the investor faces an increase in the realization-tax liability equal to the interest on the unpaid liability plus the additional tax on the asset based on a rate of return equal to the risk-free rate. Since any scheme satisfying (4) leads the investor to anticipate the same increase in tax liability (adjusting for risk) as would be imposed by the scheme in (3), intuition suggests that the potentially weaker condition (4) will also lead to holding-period neutrality.

itself is not relevant to the paper's remaining discussion. For the interested reader, an appendix to an earlier version of this paper that presents a discussion of the derivation of $V(\cdot)$ is available upon request.

PROPOSITION 1: *Condition (4) is necessary and sufficient for the achievement of holding-period neutrality for the class of assets considered in this section.*

PROOF:

At any date s , the net-of-tax value of an asset to the investor is the value of the asset A_s less the accumulated tax liability T_s . To continue to hold the asset for another instant, the investor requires a certainty-equivalent rate of return equal to the after-tax interest rate $i(1-t)$. Thus, in portfolio equilibrium:⁴

$$(5) \quad V(\dot{A}_s - \dot{T}_s) = (A_s - T_s)i(1-t).$$

Combined with equation (4), (5) implies that $V(\dot{A}_s) = iA_s$, regardless of A_s or s . Hence, (4) implies holding-period neutrality. Combined with the requirement that $V(\dot{A}_s) = iA_s$ for holding-period neutrality, (5) implies (4).

Since the certainty-equivalent value of the before-tax asset return g will equal i when an accrual-equivalent tax is imposed, it is clear that the Vickrey-type tax system described in (3) satisfies (4) and hence is holding-period neutral. As mentioned above, the challenge is to find some other tax scheme also satisfying (4) that has weaker informational requirements. One such tax system exists.

PROPOSITION 2: *Suppose the realization-tax liability at date s is*

$$(6) \quad T_s = (1 - e^{-tis}) A_s.$$

Then, the tax system satisfies (4) for all s and hence is holding-period neutral.

⁴It might be argued that the investor may not achieve an interior solution to the portfolio-choice problem in the case of assets subject to capital gains taxes. For example, one cannot freely buy and sell assets that are indexed by having already been held for a specified time period. However, the focus here is on the case in which the holding period becomes irrelevant to the portfolio-choice problem. A fortiori, the assumption of portfolio balance is justified.

PROOF:

Taking the time derivative of (6), one obtains

$$\begin{aligned}\dot{T}_s &= (1 - e^{-tis})\dot{A}_s + tie^{-tis}A_s \\ &= (1 - e^{-tis})\left(\frac{\dot{A}}{A}\right)_s A_s - (1 - e^{-tis})tiA_s + tiA_s \\ &= (1 - e^{-tis})\left[\left(\frac{\dot{A}}{A}\right)_s - ti\right]A_s + tiA_s.\end{aligned}$$

By Proposition 1, $V(\dot{A}/A) = i$ if (4) is satisfied. The strategy will be to assume $V(\dot{A}/A) = i$. Once it is proved that (4) is satisfied, the assumption will prove correct.⁵

If $V(\dot{A}/A) = i$, then $\dot{A}/A = i + \varepsilon$, where ε is a random return satisfying $V(\varepsilon) = 0$. (Note that, in general, $E(\varepsilon) \neq 0$; it is the risk premium on the risky asset). Hence,

$$\dot{T}_s = (1 - e^{-tis})(i(1 - t) + \varepsilon_s)A_s + tiA_s$$

which, by (6), may be written

$$(7) \quad \dot{T}_s = i(1 - t)T_s + tiA_s + (1 - e^{-tis})\varepsilon_s A_s.$$

Since, by construction, $V(\varepsilon) = 0$, application of $V(\cdot)$ to both sides of (7) yields (4).

A. Interpretation

Clearly, the evolution of the tax liability T_s described by (7) differs from that of the Vickrey-type system based on *ex post* returns described by (3). Since the gain $g = i + \varepsilon$, (7) differs from (3) in taxing the excess return ε at rate $(1 - e^{-tis})$ rather than t . This is a tax rate that starts at 0 and approaches 1 as s approaches ∞ . The tax rate on the excess return has no effect on the investor's welfare, however, because by construction the excess return has zero value to

him (e.g., Roger H. Gordon, 1985; Agnar Sandmo, 1985).⁶

A specific example is useful in demonstrating how this tax system works to eliminate the lock-in effect. Suppose an investor purchases an asset at some date 0 and will dispose of it with certainty at some future date s_2 . At each date s_1 between 0 and s_2 , he has the option of holding the asset or selling it for its date- s_1 value, A_1 , and buying it back. The asset's price at s_2 , A_2 , is uncertain at s_1 but is not influenced by the investor's decision.

Under the realization strategy, the investor pays a tax of $A_1(1 - e^{-its_1})$ at s_1 and $A_2(1 - e^{-it(s_2-s_1)})$ at s_2 . Under the alternative strategy, he pays $A_2(1 - e^{-its_2})$ at s_2 . A comparison of the two cases shows that the choice is between a tax payment of $e^{-its_2}(e^{its_1} - 1)A_1e^{it(s_2-s_1)}$ at s_1 versus $e^{-its_2}(e^{its_1} - 1)A_2$ at s_2 . The certainty-equivalent value of A_2 at s_1 is just $A_1e^{it(s_2-s_1)}$, however, so the investor is indifferent, *ex ante*. The two cases differ only in the *ex post* treatment of the asset's risk premium: by realizing at date s_1 , the investor prepays part of the tax that would be due at date s_2 , with the earlier payment equal in present value to the future tax it replaces.

Proposition 2 offers a very simple system of capital gains taxation. Computation of the tax burden when an asset is sold requires knowledge of the risk-free interest rate, the investor's marginal tax rate, the holding period of the asset, and the final sales price. (Nothing in the proof depends on either i or t being constant, so variations over time in rates of interest and marginal taxation present no difficulty). The initial purchase price, the pattern of accrued gains,

⁶In fact, as Gordon (1985) shows, the same general equilibrium outcome results from tax systems differing only in their treatment of excess returns, if private risk-pooling is efficient. Differences in the riskiness of after-tax returns are offset by differences in the risk characteristics of individual endowments.

⁵It is straightforward to show that this solution for required holding-period yields is unique. That is, there exists no other rate of return $j \neq i$ for which the implied tax rule corresponding to (7) is in fact consistent with the portfolio-balance condition (5) and the assumed rate of return j .

If private risk-pooling is not efficient, taxes on excess returns that have no value to investors may be pooled by the government, creating value and reducing aggregate risk. In this event, the tax rate on risk premia influences the equilibrium outcome, even though the investor's holding-period decision is not distorted.

and the asset's stochastic properties are irrelevant to the calculation. The tax itself is expressed as a time-dependent fraction of the asset's value at sale, with this fraction going from 0 at $s=0$ to 1 as s approaches ∞ .⁷

To interpret the tax formula (6), consider again the Vickrey-type tax system described in (3). For a terminal asset value of A_s , a holding period of s , and a rate of capital gain always equal to the risk-free rate (implying an initial cost of $A_s e^{-is}$), that system would impose a realization-tax liability of

$$(8) \quad T_s = \int_0^s e^{i(1-t)(s-z)} t i (A_s e^{-i(s-z)}) dz \\ = A_s (1 - e^{-its}).$$

Thus, the tax schedule (6) treats investors as if they had arrived at their current position by investing at the risk-free rate. Since in terms of certainty-equivalence this is precisely what they did, the tax system "works" in the same way that a Vickrey-type system would.⁸

Proposition 2 demonstrates that the tax system given in (6) is holding-period neutral. It is natural to ask whether there are other tax systems achieving holding-period neutrality based on the same information. However, Proposition 3 shows that the tax system already discussed is unique.

PROPOSITION 3: *The tax system described in (6) is the only one based on the information set (t, i, s, A_s) that satisfies the condition for holding-period neutrality, (4).*

⁷Since the tax liability is bounded by the asset's value, the liquidity problem is absent under this tax system. Such an accumulating tax liability over time works to remove the lock-in effect only if the tax is eventually imposed. A provision that eliminates capital gains tax liability at death, for example, might cause the lock-in effect to be exacerbated by a move to such a tax system, since investors would have an even greater incentive to hold "to the end."

⁸The admissibility of *ex ante* equivalence in the design of a tax system is discussed in Section IV.

PROOF:

Consider a tax rule based on the admissible information set:

$$(9) \quad T_s = F(t, i, s, A_s).$$

Differentiating (9) with respect to s yields

$$(10) \quad \dot{T}_s = F_s + F_A \dot{A}_s = F_s + F_A A_s (i + \varepsilon_s) \\ = F_s + F_A i A_s + F_A \varepsilon_s.$$

Applying $V(\cdot)$ to (10) and combining the result with (4) and (9) to eliminate $V(\dot{T}_s)$ and T_s , one obtains the partial differential equation

$$(11) \quad \frac{1}{i(1-t)} F_s + \frac{A_s}{1-t} F_A = F + \frac{t A_s}{1-t}.$$

Since the division of assets is arbitrary, it must be the case that F is homogeneous of degree one with respect to A_s . That is, dividing an asset into two pieces and realizing each half separately can have no effect on the capital gains tax liability. Thus, there must exist some function $F^1(\cdot)$ such that

$$(12) \quad F(i, t, s, A_s) = F^1(i, t, s) \cdot A_s.$$

Substituting the expression for F_s and F_A obtained from (12) into (11), one obtains the ordinary differential equation

$$(13) \quad \left(\frac{1}{i(1-t)} \right) \left(\frac{dF^1}{ds} \right) + \frac{1}{1-t} F^1 \\ = F^1 + \frac{t}{1-t}$$

which, combined with the initial condition $F^1(i, t, 0) = 0$, yields the unique solution $F^1(i, t, s) = (1 - e^{-its})$ and hence $T_s = F(i, t, s, A_s) = F^1(i, t, s) \cdot A_s = (1 - e^{-its}) A_s$.

The information set specified in Proposition 3 does not include the asset's initial price, though knowledge of this is required even by the current system of taxation. One's intuition might suggest that adding this piece

of information would offer an alternative rule that would also "work," similar to that given in (6) but based on the initial purchase price plus imputed interest, $A_0 e^{i's}$, rather than the sale price A_s . However, it is easy to show that this scheme would fail to satisfy condition (4). This alternative system would still encourage the holding of assets that to date had appreciated at a rate exceeding the interest rate i , since it would be imputing a normal rate of return on too low a base and, hence, would not fully eliminate the deferral advantage.

B. Extensions

One of the arguments often made for the preservation of a realization-based system of capital gains taxation is that the preferential tax treatment provided by the advantage of deferral has social value. Without judging such desirability directly, one can dispose of this argument on logical grounds by observing that the system described in (6) does not require a uniform effective tax rate on the income from all assets. A tax benefit for capital assets need not be provided via a distortionary deferral advantage.

Let t' be the tax rate on interest-bearing assets and let t be the desired effective tax rate on capital assets, perhaps lower than t' . In this case, the preceding analysis holds if one replaces the before-tax opportunity cost i with $i(1-t')/(1-t)$. That is, replacing (6) with

$$(6') \quad T_s = \left\{ 1 - \exp \left[-ti \left(\frac{1-t'}{1-t} \right) s \right] \right\} A_s$$

taxes income over time according to the rule

$$(7') \quad \dot{T}_s = i(1-t')T_s + ti \left(\frac{1-t'}{1-t} \right) A_s + \left\{ 1 - \exp \left[ti \left(\frac{1-t'}{1-t} \right) s \right] \right\} \varepsilon_s A_s.$$

Once again, the investor is charged the relevant after-tax interest rate $i(1-t')$ on the

outstanding tax liability and is taxed on the certainty-equivalent accruals of income at the capital asset's tax rate t .

Indeed, the system can be applied even if investors vary with respect to $(1-t)/(1-t')$, the ratio of their relative after-tax returns on the safe and risky assets. As long as each investor is in portfolio equilibrium, with his after-tax risk-adjusted return equal to his opportunity cost, application of the tax system in (6') implies that the investor will require a certainty-equivalent before-tax return of $i[(1-t')/(1-t)]$, even if the ratio $(1-t')/(1-t)$ varies across the population. By construction, the risk premium ε equals the total return g minus the required, risk-adjusted before-tax return $i[(1-t')/(1-t)]$, so differences in $(1-t')/(1-t)$ imply different risk premia on the same asset for different investors. However, this is precisely what gives rise to portfolio sorting and clientele formation, with investors holding diversified portfolios but gravitating toward those assets in which they obtain a relatively favorable trade-off between risk and return (Auerbach and Mervyn A. King, 1983). In equilibrium, each investor will require the available risk premium to hold each risky asset, assuming there is an interior solution to the portfolio-choice problem.⁹

III. The General Tax System

Most assets presently subject to capital gains taxes generate cash flows and are subject to tax charges before disposition of the assets themselves. In the case of corporate

⁹Such a solution will not exist, for example, if assets with different tax characteristics have the same return distributions, as in the case of perfect certainty. In such cases, constraints on investors' positions (on borrowing or short sales, perhaps) are required for any equilibrium to exist, and corner solutions for individual portfolios will arise. Here, the equivalence among after-tax returns holds only if shadow prices on the binding constraints are taken into account (see Auerbach and King, 1983). If, for example, an investor held no taxable debt, only tax-exempt municipal bonds, the appropriate after-tax opportunity cost would be the interest rate on municipal bonds.

equities, shareholders receive dividends and pay taxes on them. For other assets, taxes and cash flows may not be so closely tied. For real estate investments qualifying for accelerated depreciation allowances, for example, investors might in some years receive positive cash flows and tax refunds at the same time, and in later years they might pay taxes equal to a substantial fraction of cash flows. This section extends the previous results to the general class of assets with arbitrary patterns of cash flows and tax payments.

Let D_s be the cash distribution received at date s , and let τ_s be the tax payment made at date s . For some assets, one might impose a restriction relating τ_s to D_s , but this is unnecessary for the derivation. To the extent that there are transaction costs associated with purchasing, selling, or holding the asset, these can be treated as negative distributions.

I follow the same strategy as in Section II, first discussing the evolution of the tax liability T that is necessary to ensure holding-period neutrality. As before, I assume initially that the government wishes to tax all asset income at a single rate t .

PROPOSITION 4: *For the general class of assets just described, the following condition is necessary and sufficient for a tax to be holding-period neutral:*

$$(14) \quad V(\dot{T}_s) = i(1-t)T_s + tiA_s - \tau_s.$$

PROOF:

Following the proof of Proposition 1, note that the yield on the net of tax asset value $A - T$ must equal $i(1-t)$. This yield consists of the cash return on the asset D plus the net capital gain $\dot{A} - \dot{T}$ minus the tax payment τ ; thus,¹⁰

$$(15) \quad V(\dot{A}_s - \dot{T}_s) + D_s - \tau_s = (A_s - T_s)i(1-t).$$

Combined with equation (14), (15) implies

that $V(\dot{A}_s) + D_s = iA_s$, regardless of A_s or s . Hence (14) implies holding-period neutrality. Alternatively, combined with the requirement for holding-period neutrality that $V(\dot{A}_s) + D_s = iA_s$ (i.e., that the before-tax return required in the asset be independent of A_s or s), (15) implies (14).

Expression (14) says that, in computing their increase in tax liability \dot{T} , investors should be given credit for taxes paid currently. Again, such a provision is present in Vickrey's original scheme. As before, the rule described in (14) is less restrictive in that it applies to the valuation of returns *ex ante* rather than actual *ex post* returns in each state of nature. Once again, there is a tax system that will satisfy (14) without requiring information on the pattern of an asset's growth in value.

PROPOSITION 5: *Suppose the realization tax liability is*

$$(16) \quad T_s = (1 - e^{-tis})A_s - e^{i(1-t)s} \times \left[\int_0^s (e^{-iz} - e^{-i(1-t)z}) D_z dz + \int_0^s e^{-i(1-t)z} \tau_z dz \right].$$

Then, the tax system satisfies (14) for all s and hence is holding-period neutral.

PROOF:

Taking the time derivative of (16) and substituting the resulting expression back into (16) yields

$$\begin{aligned} \dot{T}_s &= (1 - e^{-tis})\dot{A}_s + tie^{-tis}A_s \\ &\quad + i(1-t)[T_s - (1 - e^{-tis})A_s] \\ &\quad - e^{i(1-t)s}[(e^{-is} - e^{-i(1-t)s})D_s + e^{-i(1-t)s}\tau_s] \\ &= (1 - e^{-tis})\left[\left(\frac{\dot{A}}{A}\right)_s - i\right]A_s + tiA_s \\ &\quad + i(1-t)T_s + (1 - e^{-tis})D_s - \tau_s \\ &= (1 - e^{-tis})\left[\left(\frac{\dot{A}}{A}\right)_s + D_s - i\right]A_s \\ &\quad + tiA_s + i(1-t)T_s - \tau_s. \end{aligned}$$

¹⁰The derivation assumes for simplicity that D_s and τ_s are known at date s , but this does not affect the results.

Again, without restriction (see the proof of Proposition 2) one may assume that the risk-adjusted, before-tax required return $V(\dot{A}/A) + D = i$, so that $(\dot{A}/A) + D = i + \epsilon$ with $V(\epsilon) = 0$. Thus,

$$(17) \quad \dot{T}_s = i(1-t)T_s + tiA_s - \tau_s \\ + (1 - e^{-tis})\epsilon_s A_s.$$

Since, by construction, $V(\epsilon) = 0$, application of $V(\cdot)$ to both sides of (17) yields (14).

As in the previous case, the solution involves taxing the asset's risk premium at a rate $(1 - e^{-tis})$ rather than t . A way of interpreting (16) is to rewrite it as

$$(16') \quad T_s = (1 - e^{-tis}) \left(A_s + \int_0^s e^{i(s-z)} D_z dz \right) \\ - \left(\int_0^s e^{i(s-z)} D_z dz - \int_0^s e^{i(1-t)(s-z)} D_z dz \right) \\ - \int_0^s e^{i(1-t)(s-z)} \tau_z dz.$$

The term $(A_s + \int_0^s e^{i(s-z)} D_z dz)$ is the present value, at date s , of the asset plus all previous distributions. Thus, the tax scheme begins by treating this entire value as subject to the tax rate $(1 - e^{-tis})$, as in Section II. Had all distributions been received tax-free and reinvested in the asset itself, this would be appropriate, for then the asset would be of the type analyzed there. However, because taxes have been paid in the past and the distributions invested elsewhere, two corrections are necessary for taxes already paid. The last term in (16') is a credit for taxes already paid directly on the asset, while the middle term in (16') is an imputation for taxes paid on the income generated by distributions invested in other assets facing an income tax rate t . That is, the treatment of distributions as having been reinvested in the same asset assumes that they continue to generate income at the before-tax rate of return i , adjusted for risk. Since they were actually invested in other assets, which we may assume to face an accrual-equivalent income tax rate t , we are therefore ignoring the subsequent in-

come taxes attributable to such reinvested distributions. The present value of these imputed taxes at date s is $(\int_0^s e^{i(s-z)} D_z dz - \int_0^s e^{i(1-t)(s-z)} D_z dz)$. Thus, the tax system in (16) can be interpreted as treating all distributions as being reinvested and then applying the tax scheme described in Section II but giving credit for taxes paid along the way.

Yet another interpretation of expression (16) is obtained from the following logic. As is well known, share repurchases and dividends are equivalent except for their tax treatment, and in this case, even the tax treatment is the same. Thus, one should be able to view each distribution as a share repurchase. Since each such repurchase amounts to the investor's realization of part of his assets, consistent treatment based on Proposition 1 ought to suffice. If each "partial" asset sale receives such treatment, there ought to be no deviation needed when the remainder of the asset is sold. Indeed, this conjecture is correct. Collecting terms in (16), one obtains

$$(16'') \quad T_s = (1 - e^{-tis}) A_s \\ + \int_0^s e^{i(1-t)(s-z)} [(1 - e^{-tiz}) D_z - \tau_z] dz$$

which says that the household's tax liability at date s equals the normal tax due on assets without previous distributions or tax payments plus the accumulated deficit in tax payments on previous "realizations" (i.e., distributions).¹¹

Thus, one very simple approach to the achievement of holding-period neutrality is to tax every distribution from a capital asset at the rate $(1 - e^{-tis})$, where s is the time since the asset's purchase. In this event, the informational requirements are no worse than in the previous case without distributions.

More generally, expression (16) is more complex than expression (6), but its infor-

¹¹It is particularly clear from (16'') why the initial purchase price does not appear in the tax calculation. One could view this initial cost as a negative distribution at date zero, but the appropriate tax on this negative distribution would be zero.

mational requirements are still minimal. In addition to what was needed in the previous case, the government now must also know the flows of previous taxes and distributions on the asset.

A record of previous taxes can be obtained from past tax returns. In many instances, as with common stock, the taxes are directly based on the distributions, so records of the distributions themselves are just as easily available. Even in cases for which the taxes τ and distributions D are not so simply related (real estate investments, for example), the law requires taxpayers to supply enough information so that the distributions can be calculated. For example, a real estate investor would add interest payments and depreciation deductions back to reported profits in order to calculate the distribution from a property in a given year.

As before, the tax rule can be extended to the case of different tax rates on capital assets (t) and other income (t') by replacing the interest rate i with the required before-tax return $i(1-t')/(1-t)$. For cases in which t is known, this is a simple change. There are more complicated cases, though, in which tax preferences are given not via a reduction in t but through tax credits or accelerated depreciation, each of which affects the present value of τ . In this case, it is necessary to determine what effective tax rate t is desired and to base the calculation in (16) on this value. Once this has been done, the continued presence or absence of tax credits or accelerated depreciation becomes irrelevant, for variations in these are simply offset by changes in the last term of (16).

IV. Qualifications

The system derived in the preceding sections for taxing capital gains on realization has obvious benefits, but there are potential limitations as well, some of which are discussed in this section.

A. *Ex Ante* versus *Ex Post* Taxation

One potential objection to the tax system developed in this paper is that its equivalence

to accrual taxation is on an *ex ante* basis; at each date s investors are indifferent between the increase in tax liability \dot{T}_s and accrual taxation of additional income, before they know what their income will be. However, on an *ex post* basis, the tax liabilities are not the same. In particular, it is possible for an investor to lose money continuously (A_s declining monotonically with s) and still be liable for taxes on an asset sale.

There are several responses to this criticism. First, even a system of accrual taxation, if deferred with interest as proposed by Vickrey, could lead to a positive tax liability on a capital loss.¹² Second, there are many other examples in which *ex ante* equivalence has been relied upon in the tax literature: for example, in the discussion of the conditional equivalence of consumption and wage taxes (see U.S. Treasury, 1977). Finally, the perception that this tax is unfair to those with below-normal rates of return is quite dependent on the frame of reference of a tax on *ex post* income. If, for example, one used a tax on *ex post* wealth as the frame of reference, the opposite result would hold: the tax scheme would discriminate *against* those with relatively favorable experience.

To see this, note first that a tax at rate t on an imputed rate of return i on an asset is equivalent to a wealth tax at rate ti . Thus, one may reinterpret the scheme in (6) and (16) as simulating an annual wealth tax on an asset whose value is unknown to the government. Given this interpretation, the asset whose value has risen slowly over time will have past values of wealth used for imputation [see (8)] that are too low; they will be assumed to have grown more rapidly in value over time than they actually have. The opposite will be true of assets that have appreciated rapidly.

The issue of fairness, then, involves wealth taxation to an even greater extent than

¹²For example, suppose an asset is purchased for 1 dollar, increases in value to 2 dollars and then decreases to 99 cents. The initial capital gain of 1, with interest, will exceed in absolute value the subsequent capital loss of 1.01 as long as the after-tax interest rate is greater than 1 percent.

ex ante income taxation. If the scheme considered in this paper is "unfair," then so surely must be a system of *ex post* wealth taxation. Since such property taxation is a main source of revenue for state and local governments in the United States, one must question the conclusion or at least recognize that other factors, such as ease of administrability, may outweigh the concern for *ex post* fairness in the design of policy.

B. Closely Held Assets

The system evaluated here would work best for those assets held "at arm's length," in legal terminology. This obviously includes most common stock in public corporations and other similar assets. While most common stock would be relatively easy to value and, hence, could be administered even under a system of accrual taxation, many assets in whose management the typical investor does not play an active role are nevertheless not traded at readily observable (to the government) prices. Examples would include limited partnerships and other assets which the Tax Reform Act of 1986 classifies as "passive" investments.

An asset not in this category, for example an entrepreneur-owned enterprise, would be subject to two problems. First, it would be difficult to distinguish payments to capital [taxes on which, according to (16), would be credited against ultimate capital gains tax liability] from payments to labor (which could not be so credited). Second, part of the initial value of such enterprises represents the capitalized idea of the entrepreneur. The system of retrospective taxation would tax the *income* on such initial capital appropriately but would not tax the initial income associated with the capitalization of the successful idea.¹³ This can be

compared to the current system, which taxes the initial income only upon realization and, hence, at a low effective rate, thereby introducing a powerful lock-in effect (or an incentive to be taken over by another company in order to obtain a tax-free conversion into a more diversified company's shares). In general, this is a relatively small class of assets which pose problems of administration even for the present tax system.

Just as entrepreneurs may avoid tax on labor income contributed to their enterprises under the new system, so may investors who devote labor effort to the choice of investments, in a sense producing a portfolio as the joint product of labor input and invested funds. However, this is a relatively insignificant issue for assets held at arm's length. A major exception to this conclusion would seem to arise in the case of professional securities traders, who devote most of their labor input to this endeavor. However, such income is taxed as ordinary income without any deferral advantage, even under present law. Such treatment would presumably continue even if retrospective taxation were introduced for other investors.¹⁴

V. Conclusions

This paper has presented a scheme that taxes capital gains upon realization without inducing a lock-in effect or providing the opportunity for tax arbitrage. The scheme requires information that is either publicly available (such as interest rates) or present on previous tax returns (such as past tax payments) but not the private (or potentially

case, in which the entrepreneur adds the product of his human capital and then sells the augmented asset immediately (i.e., at $s = 0$), formal adherence to the rule would produce a tax liability of zero. However, one would presumably wish to apply special rules in such special and easily identifiable cases.

¹³To see this, note that the embodiment of the idea in the asset increases the asset's value by the present value of the risk-adjusted returns that the idea is projected to yield in the future. When the investor ultimately sells the asset, the returns on the part due initially to the investor's idea are effectively taxed at the same rate as the returns on the part of the asset purchased using after-tax funds: there is no distinction regarding the source of funds, only a distinction regarding when the asset was obtained. In the simplest

¹⁴If owner-occupied housing were subject to capital gains taxation, then a significant way of achieving untaxed labor income under the new scheme (rather than having labor income eventually included in the capital gains tax base) would be to work on one's house. However, capital gains on houses are, even now, largely excluded from the tax base because of the provisions allowing the rollover of gains and the one-time exemption for individuals over age 55. Thus, even current law permits most such labor income to escape tax entirely.

even unavailable) information on the time pattern of an asset's accrued gains.

Nothing about the tax system described here requires that all asset income be taxed at the same rate for a particular investor. Purchases of certain assets can still be encouraged through a lower overall tax burden, without the need to resort to ad hoc measures such as accelerated depreciation or distortionary measures such as low rates of realization-based capital gains taxes that exacerbate the lock-in effect and the problem of tax arbitrage.

In achieving the economic benefits of accrual taxation without its associated liquidity or information problems, the new approach makes a less distortionary capital gains tax more feasible and eliminates the need for the additional distortions associated with compensating antiarbitrage provisions such as limited loss offsets.

REFERENCES

- Auerbach, Alan J., and King, Mervyn A., "Taxation, Portfolio Choice and Debt-Equity Ratios: A General Equilibrium Model," *Quarterly Journal of Economics*, November 1983, 98, 587-609.
- Constantinides, George M., "Capital Market Equilibrium with Personal Tax," *Econometrica*, May 1983, 51, 611-36.
- Gordon, Roger H., "Taxation of Corporate Capital Income: Tax Revenue versus Tax Distortions," *Quarterly Journal of Economics*, February 1985, 100, 1-27.
- Green, Jerry R. and Sheshinski, Eytan, "Optimal Capital-Gains Taxation under Limited Information," *Journal of Political Economy*, December 1978, 86, 1143-58.
- Sandmo, Agnar, "The Effects of Taxation on Savings and Risk-Taking," in A. J. Auerbach and M. Feldstein, eds., *Handbook of Public Economics*, Vol. 1, Amsterdam: North-Holland, 1985, 265-311.
- Shakow, David, "Taxation without Realization: A Proposal for Accrual Taxation," *University of Pennsylvania Law Review*, June 1986, 134, 1111-1205.
- Stiglitz, Joseph E., "The Effects of Income, Wealth and Capital Gains Taxation on Risk-Taking," *Quarterly Journal of Economics*, May 1969, 83, 203-83.
- _____, "Some Aspects of the Taxation of Capital Gains," *Journal of Public Economics*, July 1983, 21, 257-94.
- Vickrey, William, "Averaging Income for Income Tax Purposes," *Journal of Political Economy*, June 1939, 47, 379-97.
- U.S. Treasury, *Blueprints for Basic Tax Reform*, Washington, DC: U.S. Government Printing Office, 1977.

Moral Hazard and Nonmarket Institutions: Dysfunctional Crowding Out or Peer Monitoring?

By RICHARD ARNOTT AND JOSEPH E. STIGLITZ*

We examine a situation in which insurance is characterized by moral hazard. When market insurance is provided, supplementary mutual assistance between family and friends (unobservable to market insurers) will occur. When nonmarket insurers have no better information than market insurers, the mutual assistance not only crowds out market insurance but is also harmful and therefore dysfunctional. Alternatively, when nonmarket insurers can observe each other's effort perfectly, mutual assistance is beneficial. These results point to the potential importance of peer-monitoring mechanisms in mitigating moral hazard. (JEL 026)

The economics literature over the past 15 years has directed attention to the ubiquity of moral-hazard and incentives problems. One way the market responds to moral hazard is to provide only partial insurance, since then individuals still have *some* incentive to take care to avoid the accident. However, they must then bear more risk than they would like. A principal function of many nonmarket institutions, meanwhile, is to help those who have suffered some misfortune, which entails the provision of insurance: the marriage vows formalize and sanctify the mutual insurance aspects of the family; the acid test of a friend is his willingness to help in times of need; charity is regarded as meritorious and is subsidized by the government; and many government social assistance programs, such as unemployment insurance and workmen's compensation, have a strong insurance component. The importance of nonmarket insurance is illustrated by what happens if an individual

catches pneumonia as a result of going on a hiking trip with inadequate rain gear. His employer gives him compensated sick leave; part or all of his medical expenses are reimbursed by his insurance policy or the state; uncovered medical expenses may be partially deductible from his income tax; and family and friends rally round to provide other forms of support. Such extensive support, while directly helpful, deleteriously affects individuals' care to avoid accidents. In terms of the example, had the individual borne all the costs of catching pneumonia himself, he might have taken the trouble to carry adequate rain gear. Thus, it is not obvious that the insurance provided by nonmarket institutions is always beneficial or, more specifically, whether nonmarket insurance institutions, when they supplement market insurance, improve the economy's ability to handle the moral-hazard trade-off between risk-bearing and incentives.

We address this issue by inquiring whether the reciprocal provision of insurance within families and between friends, which we term nonmarket insurance,¹ is welfare-improving when it supplements market insurance. We

*Arnett: Department of Economics, Boston College, Chestnut Hill, MA 02167; Stiglitz: Department of Economics, Stanford University, Stanford, CA 94305-6072. Financial support from the National Science Foundation, the Olin Foundation, and the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged. We thank the referees and the seminar participants at Stanford University, in particular Ken Judd and Suzanne Scotchmer, for very helpful comments.

¹The term "social insurance" is perhaps more appropriate but is used in some countries to refer to social security. The term "informal insurance" is appropriate for the example, but the phenomenon we identify arises in formal, nonmarket institutions as well.

assume that a market insurer can observe his clients' market insurance purchases but not the nonmarket insurance they obtain through informal arrangements. We first show that nonmarket insurance will always be provided. Moral hazard causes fully insured individuals to expend too little effort. In response, the competitively determined market insurance contract rations the amount of insurance that can be obtained at the equilibrium price. The contract achieves this by specifying both a price and a quantity and by stipulating that it will pay in the event of accident only if the insured has no additional market insurance. Insured individuals would like to obtain additional insurance at the market price. Since they take the market insurance contract as given, they perceive that they can effectively do so by entering into informal insurance arrangements. They neglect that when everyone enters into such arrangements, the accident frequency will change, as will the market insurance contract.

After showing that nonmarket insurance will always be provided, we ask whether the level of expected utility is higher with such insurance than without it. The moral-hazard problem arises because of the inability of the insurance firm to monitor the actions of the insured. There are often other individuals, such as the members of the insured's family, who are in a better position to monitor the insured's action than the insurance firm. The welfare consequences of nonmarket insurance turn out to depend on whether these monitoring capabilities can be effectively harnessed to reduce the moral-hazard problem. If the members of the family (the providers of nonmarket insurance) do not monitor each other, we show that such nonmarket insurance always lowers welfare. The nonmarket insurance leads individuals to take less care; market insurance firms respond by providing less insurance; thus, in the new equilibrium, nonmarket insurance displaces market insurance. Since nonmarket insurance involves less risk-pooling, welfare is reduced; the less-effective insurance crowds out the more-effective insurance. The nonmarket insurance is dysfunctional.

If, alternatively, the members of the family do monitor each other, they will take

greater care than they would without monitoring, which mitigates the moral-hazard problem. Thus, there appear to be two offsetting effects, the risk-pooling advantages of market insurance versus the monitoring advantages of nonmarket insurance. However, it turns out that, with perfect monitoring within the family, the latter effect dominates, and the nonmarket insurance is welfare-enhancing.

This is an example of what we call *peer monitoring*. Peer monitoring is an important mechanism for controlling moral hazard. It arises in credit markets. In less-developed countries, loans are often made to groups of individuals; the members of a group then have an incentive to monitor each other. In developed countries, one of the functions of co-signers on loans is to provide additional monitoring. Partnership arrangements also encourage monitoring. Peer monitoring is also important in labor markets; workers are often in a better position to monitor their co-workers than are employers, which may be one of the advantages of team production.

This raises an intriguing issue related to mechanism design with a principal and many agents when moral hazard is present. The principal can monitor his agents himself (direct monitoring), hire a supervisor to monitor his agents (supervision), or set up a mechanism to induce agents to monitor each other (indirect monitoring). While the literature has considered direct monitoring (e.g., Steven Shavell, 1979; Bengt Holmstrom, 1979) and supervision (e.g., James Mirrlees, 1976; Stiglitz, 1975; Jean Tirole, 1988), it has largely ignored the design of indirect monitoring systems.² An indirect monitoring system will encourage peer monitoring through the creation of interdependence: the dependence of one agent's utility on others' effort. A good example is a university department. A faculty member's utility depends not only on his salary, perfor-

²One exception is H. Lorne Carmichael (1988). He considers peer review in the university setting, which is a form of indirect monitoring system, and argues that the institution of tenure is needed to make peer review incentive-compatible.

mance, reputation, and working conditions, but also directly on his department's and university's reputation. Furthermore, his salary, performance, reputation, and working conditions all depend to some extent on the department's quality. For these reasons, faculty members have a strong incentive to monitor one another's performance, which they do through peer review, teaching evaluations, and so on. In other contexts, production may be physically organized to facilitate peer monitoring; the open office, the assembly line, and team production are examples. In this paper, we treat the peer-monitoring system as exogenous and examine only the extreme cases in which there is either perfect monitoring or no monitoring.

I. The Basic Model without Nonmarket Insurance

Moral hazard is an asymmetric-information phenomenon, and its defining characteristic is hidden action. In the context of insurance, the probability distribution of observable outcomes depends on the insured's unobservable actions. The insurer would like to write insurance contingent on the insured's actions but, since these are unobservable, must base his insurance on observable outcomes. Because of moral hazard, there is a trade-off between risk-bearing and incentives in the provision of insurance. At one extreme, if full insurance is provided,³ the insured is equally well off whatever the outcome and therefore has no incentive to take care. At the other extreme, if no insurance is provided, the individual faces the appropriate incentives but is fully exposed to risk.

We first describe the canonical moral-hazard model without nonmarket insurance (Arnott and Stiglitz, 1988a). There is a single, fixed-damage accident. The probability of its occurrence, p , depends on the individual's effort at accident avoidance, e . The probability-of-accident function is strictly convex: $p' < 0$, $p'' > 0$. The individual's

wealth is w , and d is the damage caused by the accident.

If an accident occurs, the individual receives a (net of premium) payout of α , and his consumption is

$$(1a) \quad y_1 = w - d + \alpha.$$

If an accident does not occur, he pays the premium β , and his consumption is

$$(1b) \quad y_0 = w - \beta.$$

For simplicity, we assume a separable, event-independent utility function.⁴ Expected utility is then

$$(2) \quad EU = u(y_0)(1 - p) + u(y_1)p - e$$

with $u' > 0$, $u'' < 0$.

The individual chooses effort so as to maximize expected utility, taking α and β as given:⁵

$$(3) \quad e(\alpha, \beta) = \underset{(e)}{\operatorname{argmax}} EU(\alpha, \beta, e).$$

It is straightforward to show that, with the form of utility function assumed, $\partial e / \partial \alpha < 0$ and $\partial e / \partial \beta < 0$; as more insurance is provided, the individual reduces effort. Substitution of (3) into (2) gives $V(\alpha, \beta) \equiv EU(\alpha, \beta, e(\alpha, \beta))$, from which an individual's indifference curves in $\alpha - \beta$ space can be plotted (see Fig. 1). The slope of an indifference curve is

$$(4) \quad \left. \frac{d\beta}{d\alpha} \right|_{\bar{V}} = \frac{u'_1 p}{u'_0(1 - p)}$$

where $u_0 \equiv u(y_0)$ and $u_1 \equiv u(y_1)$. The locus of (α, β) for which there are zero profits, the zero-profit locus (ZPL), is $(1 - p)\beta - p\alpha = 0$. Substitution of (3) into this equation

⁴The general form of the utility function with event i is $U_i(y_i, e)$. Separability implies $U_i(y_i, e) = u_i(y_i) - e$. Event-independence implies that the utility-from-consumption function $u_i(y)$ is independent of the event; the accident causes neither pain nor pleasure and does not alter tastes.

⁵Throughout the paper, we ignore the complications that arise from the possibility that $e = 0$.

³Full insurance—equalization of the marginal utilities of income in all events—implies no incentive to take care when the utility function is event-independent, which we assume, but not generally otherwise.

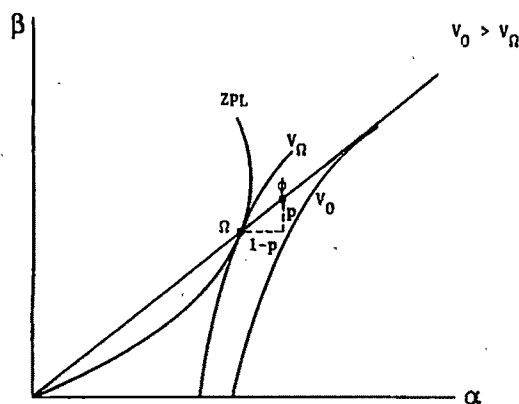


FIGURE 1. THE INDIVIDUAL PERCEIVES THAT HE CAN OBTAIN ADDITIONAL INSURANCE AT THE MARKET PRICE BY ENTERING INTO A MUTUAL INSURANCE PACT

gives the ZPL as a function of α and β , which is plotted in Figure 1. Its slope is

$$(5) \quad \left. \frac{d\beta}{d\alpha} \right|_{ZPL} = \frac{p + (\alpha + \beta)p'(\partial e / \partial \alpha)}{(1-p) - (\alpha + \beta)p'(\partial e / \partial \beta)}$$

In Figure 1, Ω is the point of optimal insurance, contingent on the unobservability of effort;⁶ it occurs at the point of maximum utility on the zero-profit locus.

We now investigate competitive equilibrium for this model with insurance purchases observable (Arnott and Stiglitz, 1988a, b). Define $q \equiv \beta/\alpha$ to be the price of insurance. Since Ω is on the ZPL, $q_\Omega \equiv (\beta/\alpha)_\Omega = [p/(1-p)]_\Omega$. The precise shapes of the indifference curves and the ZPL are inessential for our analysis.⁷ What is essen-

⁶ Ω is the optimal *deterministic* contract. The circumstances in which randomization is Pareto-improving are investigated in Arnott and Stiglitz (1988c).

⁷Indifference curves are positively sloped and can be nonconvex

$$\left(\left. \frac{d^2\beta}{d\alpha^2} \right|_v > 0 \right)$$

as can be seen from (4). As the amount of insurance provided increases, holding utility fixed, u'_1/u'_0 falls,

tial, however, is that at Ω the slope of the ZPL exceeds the price of insurance. Since indifference curves are positively sloped and since Ω is at a point of tangency of an indifference curve and the ZPL, the ZPL must be positively sloped at Ω . Now consider increasing α by one unit. To maintain zero profits, β must be increased by more than the price of insurance, $q_\Omega = [p/(1-p)]_\Omega$. An increase in β of $[p/(1-p)]_\Omega$ causes zero profit to be made on the *marginal* unit of insurance; but the increase in α raises the probability of accident, and so β has to be increased further to offset losses on *inframarginal* units of insurance. Hence,

$$(6) \quad \left(\left. \frac{d\beta}{d\alpha} \right|_v \right)_\Omega = \left(\left. \frac{d\beta}{d\alpha} \right|_{ZPL} \right)_\Omega > q_\Omega$$

Since at Ω the slope of the indifference curve exceeds the price of insurance, individuals would like to acquire more insurance at the price q_Ω (see Pauly, 1974). Thus, competitive decentralization of Ω requires rationing of insurance at the price q_Ω .⁸ The intuition for this result is that, because of moral hazard, at any price of insurance, the individual expends too little effort. Rationing induces him to increase his effort. The easiest⁹ way for such rationing to be

but since more insurance causes the probability of accidents to rise, $p/(1-p)$ increases. The zero-profit locus is positively sloped for small amounts of insurance but may bend backwards.

⁸The point Ω can be decentralized by the insurer offering any locus of (α, β) such that Ω is the point of maximum utility on the locus and insisting that he be the sole insurer. One such possibility entails the insurer offering the zero-profit locus. This particular form of nonlinear pricing in this context was considered by Elhanan Helpman and Jean-Jacques Laffont (1975). The analysis of the paper carries through whatever form of pricing is employed to decentralize Ω .

⁹Insurers could alternatively agree to sell a client who has purchased a total amount of insurance from other insurers of (α', β') an insurance policy $(\hat{\alpha}, \hat{\beta}) = (\alpha_\Omega - \alpha', \beta_\Omega - \beta')$, which would bring his total insurance up to $(\alpha_\Omega, \beta_\Omega)$. However, this is unnecessarily complicated. In fact, almost all insurance policies contain exclusivity provisions (life and air-flight accident insurance are the exceptions, but moral hazard is unimportant for both types of insurance).

accomplished is for each insurer to offer the contract $(\alpha_\Omega, \beta_\Omega)$, conditional on its clients purchasing no additional insurance. This condition, which we term the exclusivity provision, is enforceable, since it is assumed that insurers are able to observe all insurance purchases.

To sum up, with moral hazard, when non-market insurance is absent and insurance purchases are observable, the social optimum conditional on the unobservability of effort is decentralizable. Competitive equilibrium entails each individual being restricted to purchase all his insurance from a single insurer and being rationed in the amount of insurance he can purchase at the equilibrium price.

II. Effort Unobservable by the Nonmarket Insurer

We now consider the simplest possible extension of this model that allows for the simultaneous provision of market and non-market insurance. We assume that, although an insurer can observe his clients' insurance purchases, he cannot observe the nonmarket insurance they acquire.¹⁰

As described earlier, nonmarket insurance may be provided through many institutions. To simplify, we treat only nonmarket insurance provided reciprocally by pairs of symmetrical individuals. We label the two partners in a pair, H (husband) and W (wife). The two partners have the same tastes and probability-of-accident functions, and their accident probabilities are statistically independent.¹¹ The form of the non-market insurance is as follows: H and W agree that if one spouse has an accident and

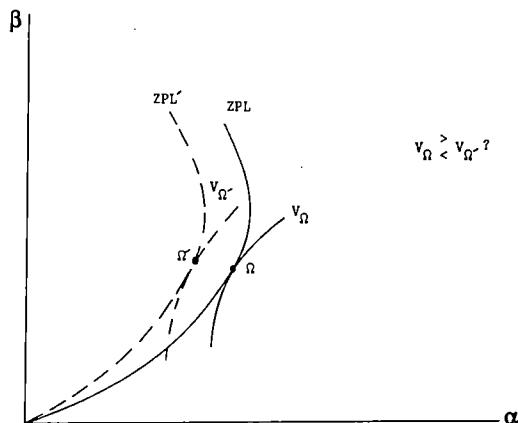


FIGURE 2. THE EFFECT OF NONMARKET INSURANCE WHEN IT REDUCES EFFORT

the other does not, the latter will transfer δ to the former.

The issues to be addressed can be posed in terms of Figure 2. The solid-lined ZPL and V_Ω are the zero-profit locus and equilibrium-indifference curve in the absence of nonmarket insurance (the same as in Fig. 1); the dashed-lined ZPL and V_Ω' are the corresponding curves with nonmarket insurance. Assume, for the sake of argument, that the provision of nonmarket insurance reduces effort. There are then two offsetting effects on equilibrium utility. On one hand, with α and β held fixed, the provision of nonmarket insurance increases utility; it shifts indifference curves to the left. On the other hand, since by assumption the provision of nonmarket insurance increases the probability of accident, the insurance firm must lower the payout for any level of the premium to continue making zero profits; as a result, the ZPL also shifts to the left. The analysis that follows examines which effect is dominant.

In this section, we characterize the equilibrium for the case in which an individual's accident-avoidance effort is observable by neither his partner nor market insurers. Equilibrium may entail a combination of market and nonmarket insurance. Subsequently, we shall investigate the efficiency properties of the equilibrium.

¹⁰If an insurer were able to observe the nonmarket insurance that his clients acquire, he could write the contract contingent on their nonmarket insurance. The resulting equilibrium would be efficient conditional on the information available (the efficiency result depends on there being only one commodity and one type of accident; see Arnott and Stiglitz, 1986, 1989).

¹¹This assumption simplifies the analysis and does not affect the qualitative results. Depending on context, the accident probabilities of the partners may be positively correlated.

There are four events: 1) neither the individual nor his partner has an accident; 2) the individual has an accident, but his partner does not; 3) only the partner has an accident; 4) both have accidents. Let e denote the individual's effort and \bar{e} his partner's. Then, the probability that neither the individual nor his partner has an accident is $[1 - p(e)][1 - p(\bar{e})]$ and similarly for the other events. We assume that the market insurer sells individual rather than group insurance policies¹² (i.e., he sells each partner a policy (α, β) ; if neither individual has an accident, each has to pay the market insurer the premium β , etc.).

Thus, an individual's expected utility is

$$(7) \quad EU = u(w - \beta)[1 - p(e)][1 - p(\bar{e})] \\ + u(w - d + \alpha)p(e)p(\bar{e}) \\ + u(w - \beta - \delta)[1 - p(e)]p(\bar{e}) \\ + u(w - d + \alpha + \delta)p(e) \\ \times [1 - p(\bar{e})] - e$$

which may be written more succinctly as

$$(7') \quad EU = u_0(1 - p)(1 - \bar{p}) + u_1 p \bar{p} \\ + u_2(1 - p)\bar{p} + u_3 p(1 - \bar{p}) - e$$

where $u_0 \equiv u(w - \beta)$, $u_1 \equiv u(w - d + \alpha)$, $u_2 \equiv u(w - \beta - \delta)$, $u_3 \equiv u(w - d + \alpha + \delta)$, and $\bar{p} \equiv p(\bar{e})$.

We assume that H and W are smart and take into account how the other will adjust effort in response to a change in δ .¹³ Both

assume that the market contract will be unaffected by their actions, which is perfectly reasonable in the atomistic environment we envisage.

We adopt the Nash assumption. W, in deciding on her level of effort, takes α , β , and δ as well as H's effort as fixed. W believes that H is rational and selfish and will accordingly choose the level of effort that maximizes his expected utility, given that W is acting similarly.¹⁴ Then, from (7), the equation characterizing her level of precaution is

$$(8) \quad [-u_0(1 - \bar{p}) + u_1 \bar{p} - u_2 \bar{p} \\ + u_3(1 - \bar{p})]p' - 1 = 0$$

which gives

$$(9a) \quad e = \hat{e}(\alpha, \beta, \delta, \bar{p})$$

and by symmetry

$$(9b) \quad \bar{e} = \hat{e}(\alpha, \beta, \delta, p).$$

Combining (9a) and (9b) yields

$$(10a) \quad e = e(\alpha, \beta, \delta)$$

$$(10b) \quad \bar{e} = e(\alpha, \beta, \delta).$$

From (7), the individual and his partner perceive expected utility to be related to δ in the following way:

$$(11) \quad \frac{\partial EU}{\partial \delta} = [-u_2'(1 - p)\bar{p} + u_3'p(1 - \bar{p})] \\ + \{[-u_0'(1 - \bar{p}) + u_1'\bar{p} - u_2'\bar{p} \\ + u_3'(1 - \bar{p})]p' - 1\} \frac{\partial e}{\partial \delta} \\ + [-u_0'(1 - p) + u_1'p \\ + u_2'(1 - p) - u_3'p] \bar{p}' \frac{\partial \bar{e}}{\partial \delta}.$$

¹²Where the insurance company cannot identify the nonmarket insurance partners (friends, for example), this is the natural assumption. However, in other contexts, notably the family, the market typically provides group policies. We have run through the analysis when the market provides group rather than individual policies. Group policies are Pareto-superior, since they contain an extra policy parameter: an insurance policy specifies the group premium payable if neither has an accident, the net payout if only one has an accident, and the net payout if both have an accident. In other respects, the qualitative results are the same when the market provides group policies as when it provides individual policies.

¹³We would obtain the same qualitative results if we assumed instead that H and W ignore that the other will adjust effort in response to a change in δ .

¹⁴This is not an infinitely repeated game. If it were, the cooperative outcome might be obtained. We comment on this later.

In so doing, they neglect that, since other couples too behave in this way, insurance companies adjust α and β in response to a change in δ . Combining (8), (10a), (10b), and (11), and noting that the equilibrium is symmetric, gives

$$(12) \quad \frac{\partial EU}{\partial \delta} = (-u'_2 + u'_3)(1-p)p + [1 + (u_2 - u_3)p'] \frac{\partial e}{\partial \delta}.$$

Furthermore, from (8),

$$(13) \quad \frac{\partial e}{\partial \delta} = - \frac{[u'_2 p + u'_3(1-p)]p'}{(p''/p') + (p')^2(u_0 + u_1 - u_2 - u_3)} < 0.$$

At Ω (the competitive equilibrium in the absence of nonmarket insurance), $\delta = 0$, $1 + (u_2 - u_3)p' = 0$ [from (8) since $u_0 = u_2$ and $u_1 = u_3$], and $-u'_2 + u'_3 > 0$ (incomplete insurance). Hence, from (12),

$$(14) \quad \left. \frac{\partial EU}{\partial \delta} \right|_{\Omega} = (-u'_2 + u'_3)(1-p)p > 0.$$

Thus, at the competitive equilibrium in the absence of nonmarket insurance, the partners perceive a mutual insurance pact to be beneficial and would therefore provide one another with nonmarket insurance to supplement their market insurance. The intuition for this result is as follows. At Ω , the partners are rationed in the amount of insurance they can purchase at the price q_{Ω} . Each perceives that, by entering into a mutual insurance pact, he can acquire additional insurance at this price, which pays out when he, but not his partner, suffers an accident. More specifically, at Ω , since

$$\frac{\partial EU}{\partial \alpha} = pu'_3 \quad \text{and} \quad \frac{\partial EU}{\partial \beta} = -(1-p)u'_2$$

from (12),

$$(15) \quad \frac{\partial EU}{\partial \delta} = (1-p) \frac{\partial EU}{\partial \alpha} + p \frac{\partial EU}{\partial \beta}.$$

An individual regards a unit increase in δ as equivalent to a unit increase in α with probability $1-p$ (the probability that his partner is not sick when he is) combined with a unit increase in β with probability p (the probability that his partner is sick when he is not), or equivalently, as an expected increase of $1-p$ in the amount of insurance obtained at the price q (i.e., movement from Ω to ϕ in Fig. 1). As already noted, in reasoning in this way, individuals neglect that, when everyone enters into such a pact, which reduces effort [eq. (13)] and increases the probability of accident, market insurers are forced to offer a less attractive contract in order to maintain zero profits.

H and W choose δ to maximize their expected utilities, taking α and β as given. From $\partial EU / \partial \delta = 0$, $e = \bar{e}$, and (8), one obtains $\delta = \delta(\alpha, \beta)$. By observing how the probability of accident responds to changes in α and β , market insurers will implicitly take into account that δ responds to α and β according to $\delta = \delta(\alpha, \beta)$. Competition, meanwhile, will continue to result in the equilibrium market contract maximizing expected utility subject to zero profits. Thus, in the presence of nonmarket insurance, the equilibrium market contract maximizes

$$(16) \quad \begin{aligned} EU = & u(w - \beta)(1-p)^2 \\ & + u(w - d + \alpha)p^2 \\ & + u(w - \beta - \delta)(1-p)p \\ & + u(w - d + \alpha + \delta)p(1-p) - e \end{aligned}$$

subject to

$$\begin{aligned} (i) \quad & \beta(1-p) - \alpha p = 0 \\ (ii) \quad & e = e(\alpha, \beta, \delta(\alpha, \beta)) \end{aligned}$$

where (ii) is obtained by combining (10a) and $\delta = \delta(\alpha, \beta)$.

Given the assumed information technology, it can be shown that the nonmarket insurance is unambiguously harmful and dysfunctional. The line of proof is straightforward: welfare is at least as high if the market insurer chooses α , β , and δ as if he chooses just α and β , with δ being chosen by the nonmarket insurer; and if the market

insurer chooses α , β , and δ , he will set $\delta = 0$.

The equilibrium without nonmarket insurance cannot be improved upon, and if it were possible, it would be desirable to outlaw the provision of nonmarket insurance. The intuitive rationale for this result is as follows. The provision of nonmarket insurance does not enhance the risk-sharing capabilities of the economy. Rather, such insurance crowds out market insurance. Not only is it less effective than market insurance since it randomizes an individual's event-contingent consumption and is provided by a risk-averse agent (see John M. Marshall, 1976), but also the simultaneous provision of market and nonmarket insurance violates exclusivity, which typically creates uninternalized externalities (see Arnott and Stiglitz, 1989).

The above analysis was predicated on the assumptions that a market which provides insurance against the accident in question exists and that there are no transaction costs associated with the provision of insurance. If market insurance against a given accident does not in fact exist, voluntary nonmarket insurance is unambiguously beneficial.¹⁵ When transactions costs are present, nonmarket insurance may be beneficial if it is provided at lower transaction cost than market insurance.

III. Effort Observable by the Nonmarket Insurer

This case is more interesting, since there appear to be two offsetting effects. On one hand, because individuals have information on their partner's effort, which an insurance company does not, the provision of nonmarket insurance has the potential of enhancing the risk-sharing capabilities of the economy. On the other hand, the provision of insurance by a risk-neutral agent is typically more efficient than provision by a risk-averse

agent, if they have access to the same information. Furthermore, the simultaneous provision of market and nonmarket insurance violates exclusivity. This line of reasoning suggests that the provision of nonmarket insurance in this case may be beneficial in some circumstances and harmful in others.

We continue with the same model. When effort is observable within the family but not to the insurance firm, and when, as we have assumed, individuals are identical, family members will effectively choose the level of precaution to take cooperatively. Each will take α and β to be fixed and choose δ and e to maximize

$$(17) \quad EU = u_0(1-p)^2 + u_1p^2 + u_2(1-p)p + u_3(1-p)p - e.$$

This yields the following first-order conditions:

$$(18a) \quad e: [-2(1-p)u_0 + 2pu_1 + (1-2p)(u_2 + u_3)]p' = 1$$

$$(18b) \quad \delta: (-u'_2 + u'_3)p(1-p) = 0.$$

Equation (18b) implies that

$$(18b') \quad \delta = \frac{d - \alpha - \beta}{2}.$$

Because the partners can observe each other's effort and treat α and β as fixed, they perceive there to be no moral-hazard problem associated with the insurance they provide and hence provide full insurance (or as full as possible). This stands in contrast to the previous section where, as a result of the inability of each partner to observe the other's effort, only partial nonmarket insurance was provided [see (12)].

The insurance firm effectively chooses α and β to maximize expected utility, subject to (18a), (18b), and the zero-profit constraint. The competitive equilibrium with nonmarket insurance is characterized by the constraints and first-order conditions of this program.

¹⁵Market insurance is generally unavailable for the multitudinous small (but cumulatively substantial) risks faced in everyday life. How nonmarket institutions handle such risks is an important and interesting question.

We now investigate the welfare properties of the equilibrium. To do this, we assume that the planner chooses α , β , and δ , knowing that individuals choose e according to (18a), which takes account of the fact that δ is chosen with effort observable, and subject to the break-even constraint on market insurance.

Substituting the zero-profit constraint into (17) gives

$$(19) \quad EU(\beta, \delta) = u(w - \beta)(1 - p)^2 \\ + u\left(w - d + \frac{\beta(1 - p)}{p}\right)p^2 \\ + u(w - \beta - \delta)p(1 - p) \\ + u\left(w - d + \frac{\beta(1 - p)}{p} + \delta\right) \\ \times p(1 - p) - e.$$

The corresponding first-order condition for δ [using (18a)] is

$$(20) \quad \frac{\partial EU}{\partial \delta} = (-u'_2 + u'_3)p(1 - p) \\ - \left(\beta u'_1 p' + \frac{\beta(1 - p)}{p} u'_3 p'\right) \frac{\partial e}{\partial \delta} = 0.$$

Using (18a) again,

$$(21) \quad \frac{\partial e}{\partial \delta} = \frac{(1 - 2p)(u'_3 - u'_2)}{\Delta}$$

where

$$\Delta = \frac{\beta p'}{p^2} [2pu'_1 + (1 - 2p)u'_3] \\ - 2p'(u_0 + u_1 - u_2 - u_3) - \frac{p''}{(p')^2}.$$

Substituting (21) into (20) gives

$$(20') \quad \frac{\partial EU}{\partial \delta} = \frac{(u'_3 - u'_2)}{\Delta} \\ \times \left[\beta u'_1 p' - 2p'p(1 - p) \right. \\ \times (u_0 + u_1 - u_2 - u_3) \\ \left. - \frac{p''}{(p')^2} p(1 - p) \right].$$

Both Δ and the expression in brackets are unambiguously negative, and hence $\partial EU / \partial \delta = 0$ if and only if $u'_3 = u'_2$ [i.e., iff $\delta = \delta^* \equiv (d - \alpha - \beta)/2$]. Furthermore, $u'_3 > u'_2$ for $\delta < \delta^*$ and $u'_3 < u'_2$ for $\delta > \delta^*$, and so δ^* is the utility-maximizing δ . Thus, when effort is observable by the nonmarket insurer, the equilibrium is constrained efficient.

Does the provision of nonmarket insurance in this case stimulate or discourage effort? To answer this, we hold α and β fixed and increase δ from 0 to $(d - \alpha - \beta)/2$. Though the increase in δ unambiguously reduces risk, it has an ambiguous effect on effort. The risk reduction, by itself, encourages a reduction in effort. However, as δ increases, individuals become less selfish in their choice of effort. From (21), the former effect dominates if $p < 1/2$, and the latter effect otherwise. In the normal case, $p < 1/2$, the trade-off illustrated in Figure 2 is present, but the direct utility-increasing effect of nonmarket insurance unambiguously dominates the effort-reducing effect.

We can draw together the results in the following inequalities:

$$(22) \quad EU^1 > EU^{NMO} > EU^M > EU^{NMU}$$

where EU^1 is expected utility in the first-best case (with effort observable to both market and nonmarket insurers), EU^{NMO} is expected utility with nonmarket insurance and with effort observable to the nonmarket insurer but not the market insurer, EU^M is expected utility with only market insurance

and with effort unobservable to the market insurer, and EU^{NMU} is expected utility with nonmarket insurance and with effort unobservable to both market and nonmarket insurer.

IV. Discussion

The above models were rather stark. Some discussion and interpretation will therefore be useful. The results of the two cases analyzed above lead naturally to the conjecture that, in intermediate situations in which nonmarket insurers observe their partners' effort imperfectly but better than the market insurer, an excessive amount of nonmarket insurance will be provided which may or may not be better than no nonmarket insurance at all. The analysis could be extended to compare the optimal and equilibrium numbers of members in a nonmarket insurance group; in a large group, there is greater diversification of risk but more imperfect observability.

We distinguished between the two cases treated on the basis of the observability of one partner's effort by the other. The essential difference between the two cases was, however, the severity of moral hazard within the partnership, and this depends on more than the observability of effort. Such factors as the duration of the partnership, the discount rate, the frequency of accidents, the severity of punishment for renegeing on an agreement, the power of reputation and social pressure, and nonmonetary rewards from cooperation, are also important.¹⁶

We provided a rather narrow interpretation of our model. Other interpretations are possible. A market insurer and his clients

can be replaced by a principal and his agents, and the partnerships (with some elaboration of the model) can be replaced by secondary markets.¹⁷

In the above analysis, we took the observability of one partner's accident-prevention effort by the other as exogenous and considered only the extreme cases of unobservability and perfect observability. As we noted earlier, the degree of observability depends on the indirect monitoring system—the means by which one partner observes the other's effort, as well as the incentives to do so. Furthermore, in a fuller analysis, the indirect monitoring system would be treated as endogenous. Indeed, how indirect monitoring systems develop in nonmarket insurance institutions is an exciting research topic, as is the more general issue of how principals should design indirect monitoring systems so as to mitigate the incentives problem.

There is widespread belief that when significant market failure occurs, there are strong incentives for nonmarket institutions to develop which go at least part of the way toward remedying the deficiency.¹⁸ This paper has provided a counterexample¹⁹ in

¹⁷One can develop a typology in terms of the primary and secondary insurance arrangements, each of which may be market or nonmarket. We have considered the case in which the primary insurance arrangement is a market and the secondary arrangement is a partnership. The case primary-market/secondary-market is an insurance market in which exclusivity cannot be enforced.

¹⁸In anthropology there is a functionalist tradition of long standing which attempts to explain social institutions (political, economic, sociological, cultural, and psychological) as functional adaptations to a society's environment or ecosystem. Functionalist theories differ in their degree of subtlety and sophistication and in their emphasis, but none seems to make a sharp distinction between equilibrium and optimum. In most theories, however, there seems to be a presumption that institutional adaptation to the environment is efficient. See Roger Keesing (1981) for an informative discussion of contemporary traditions in anthropology.

¹⁹In our example, with effort unobservable to nonmarket insurers, the market by itself is constrained Pareto-efficient. However, there is *perceived* market failure, and the nonmarket institution (the provision of supplementary nonmarket insurance) arises in response to this perceived market failure.

¹⁶Altruism is also a factor in nonanonymous relationships. In terms of the model, let EU be an individual's expected utility and \widetilde{EU} his partner's. The individual maximizes welfare $W = EU + \lambda \widetilde{EU}$, where $\lambda = 0$ corresponds to the selfish case, $\lambda = 1$ to balanced altruism, and $\lambda = \infty$ to selfless altruism. If both partners are equally altruistic and cooperate (the effort-observable case), the outcome is independent of the degree of altruism. If both partners are equally altruistic and do not cooperate (the effort unobservable case), effort increases with the degree of altruism.

which a nonmarket institution arises spontaneously (through the uncoordinated actions of atomistic agents), which is completely dysfunctional (has effects opposite to those intended).²⁰ In our stark model, though the market response to imperfect information (the rationing of insurance) did indeed give rise to a nonmarket response (nonmarket insurance), whether the nonmarket response was welfare-enhancing turned out to depend on whether the nonmarket institution was informationally advantaged relative to the market institution. Our example illustrates, in a vivid way, the functionalist fallacy: the fact that an institution (nonmarket insurance) has a clearly identifiable function (to improve risk-sharing by supplementing the rationed insurance provided by the market) does not mean, within a general equilibrium context, that it actually performs that function (indeed, in one case, the nonmarket insurance was completely dysfunctional). We speculate that the possible dysfunctionality of nonmarket institutions is a general phenomenon. This, too, is an interesting topic for future research.

In an expanded version of our model in which there are many kinds of accidents and many commodities, the market is not constrained Pareto-efficient; there is genuine potential market failure (Arnott and Stiglitz, 1989). Our result concerning the possible dysfunctionality of spontaneous nonmarket institutions carries over to this more realistic setting.

²⁰In one sense, this result should come as no surprise, since it is by now well recognized that, even in large economies, Nash equilibria are Pareto-efficient only under special circumstances. One of the great achievements of modern economics was to identify a special set of assumptions under which competitive economies are Pareto-efficient.

George Akerlof (1980) has argued that inefficient social customs may persist as Nash equilibria and that there can be an arbitrarily large set of social customs sustainable as Nash equilibria. The point in our paper is related but different. Akerlof considers the possible persistence of inefficient institutions but does not investigate how the institutions came into being. We show not only that an inefficient institution can persist, but also that it can arise spontaneously. Furthermore, while in the Akerlof model there are multiple equilibria of which some may be efficient, in our model there is a unique equilibrium.

V. Conclusion

In this paper we developed a simple moral-hazard model in which individuals have an incentive to supplement market insurance with nonmarket insurance that is unobservable to the market insurer. Whether this nonmarket insurance is socially beneficial depends on whether the nonmarket insurance partners can monitor each other's accident-prevention effort better than the market. In the extreme case in which the partners have no more information than the market, the nonmarket insurance is unambiguously harmful. The nonmarket insurance crowds out the market insurance and results in less risk-spreading; it is therefore completely dysfunctional. In the other extreme case, in which the partners can observe each other's accident-prevention effort perfectly, the nonmarket insurance is ameliorative, and the equilibrium is constrained efficient. The peer monitoring by the nonmarket insurers is effectively utilized to mitigate the moral hazard and to improve the risk-sharing capabilities of the economy.

The simple model raises two broad questions which go beyond the context of moral hazard in insurance markets. In what other situations may nonmarket institutions be dysfunctional? And how may the economy utilize peer-monitoring systems to improve efficiency?

REFERENCES

- Akerlof, George, "A Theory of Social Custom, of which Unemployment May Be One Consequence," *Quarterly Journal of Economics*, June 1980, 94, 749-75.
- Arnott, Richard and Stiglitz, Joseph E., "Moral Hazard and Optimal Commodity Taxation," *Journal of Public Economics*, February 1986, 29, 1-24.
- _____, (1988a) "The Basic Analytics of Moral Hazard," *Scandinavian Journal of Economics*, 1988, 90 (3), 383-413.
- _____, (1988b) "Equilibrium in Competitive Insurance Markets with Moral Hazard," mimeo, Stanford University, 1988.

- _____, (1988c) "Randomization with Asymmetric Information," *Rand Journal of Economics*, Autumn 1988, 19, 344-62.
- _____, "The Welfare Economics of Moral Hazard," in H. Loubergé, ed., *Information and Insurance: Essays in Memory of Karl H. Borch*, Norwell, MA: Kluwer, 1989.
- Carmichael, H. Lorne, "Incentives in Academics: Why is There Tenure?" *Journal of Political Economy*, June 1988, 96, 453-72.
- Helpman, Elhanan and Laffont, Jean-Jacques, "On Moral Hazard in General Equilibrium Theory" *Journal of Economic Theory*, February 1975, 10, 8-23.
- Holmstrom, Bengt, "Moral Hazard and Observability," *Bell Journal of Economics*, Spring 1979, 10, 74-91.
- Keesing, Roger M., *Cultural Anthropology: A Contemporary Perspective*, 2nd Ed., New York: Holt, Rinehart, and Winston, 1981.
- Marshall, John M., "Moral Hazard," *American Economic Review*, December 1976, 66, 880-90.
- Mirrlees, James A., "The Optimal Structure of Incentives and Authority within an Organization," *Bell Journal of Economics*, Spring 1976, 7, 105-31.
- Pauly, Mark, "Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection," *Quarterly Journal of Economics*, February 1974, 88, 44-54.
- Shavell, Steven, "On Moral Hazard and Insurance," *Quarterly Journal of Economics*, November 1979, 93, 541-62.
- Stiglitz, Joseph E., "Incentives, Risk and Information: Notes Towards a Theory of Hierarchy," *Bell Journal of Economics*, Autumn 1975, 6, 552-72.
- Tirole, Jean, "The Multicontract Organization," *Canadian Journal of Economics*, August 1988, 21, 459-66.

A Theory of Inefficient Intrafirm Transactions

By JULIO J. ROTEMBERG*

I consider a model in which the threat of customer departures induces sellers to supply high-quality goods. Permanent attachment of buyer and seller such that transactions take place inside a firm raises the social cost of delivering high quality. Yet, such costly integration is often profitable, because prices exceed marginal cost at equilibria where market transactions provide high quality. This theory can rationalize the empirical finding that middle managers are averse to transactions between profit centers. (JEL 611)

Unrelated individuals work together within firms where relationships of authority mold individual behavior. At least since Ronald Coase (1937), economists have viewed the abandonment of the market implied by these relationships as a socially desirable adaptation.¹ Oliver Williamson (1985 p. 17), for instance, states: "This book advances the proposition that the economic institutions of capitalism have the main purpose of economizing on transactions costs." In this view, the routing of transactions inside firms raises profits together with social welfare.

In this paper, I follow Coase (1937) in that I assume that integration and the concomitant abandonment of market transactions occur only when this is privately profitable. However, because product market imperfections exist, their private profitability is consistent with lack of social desirability. When goods markets are distorted, firms generally earn rents. Thus, firms would willingly tolerate organizational inefficiencies if such inefficiencies help them capture these rents. Some of the economy's rents may be

thus dissipated, and this may help explain why conventional measurements of profits do not uncover much monopolistic conduct.

I focus on a situation in which it is difficult to ensure that the good bought by one agent from another is of high quality. Quality is something the buyer can recognize *ex post* but not something that can be stipulated in advance when the agents enter into contracts. As shown by Benjamin Klein and Keith Leffler (1981) spot markets deal well with this problem, at least when price exceeds marginal cost. Then, sellers in spot markets try hard to provide high quality for fear that customers will change suppliers whenever they have an unsatisfactory experience. Such fear only arises if the customer's future business leads to profits and if it is reasonable to suppose that the customer will indeed leave when he is dissatisfied.

Even when suppliers try hard to provide high-quality goods, the good will sometimes fall short of the customer's expectation. This is particularly true when quality refers not to a technical characteristic, but to the way the good is delivered. An example of this is the supplier's ability to alter delivery schedules and specifications to accommodate changes in the customer's needs.

As in Franklin Allen and Gerald Faulhaber (1988), quality can fall below expectation for two reasons. The first is temporary bad luck which makes it impossible for the supplier to accommodate the buyer in this particular instance. Such temporary bad luck bears no relationship to the likelihood that

*Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139. I thank three anonymous referees, Oliver Hart, Jim Poterba, Jean Tirole, Garth Saloner, and participants at seminars at Boston University, M.I.T., the University of Chicago, and Yale University for comments and the NSF and Sloan Foundations for research support.

¹See Armen Alchian and Harold Demsetz (1972) and Sanford Grossman and Oliver Hart (1986), as well as the references in Bengt Holmstrom and Jean Tirole (1989).

the seller will offer high quality in the future. The second reason is that the supplier is sometimes genuinely incompetent. Then, he is likely to continue delivering bad quality in the future. A dissatisfied customer will typically not know whether he has faced bad luck or incompetence. Fearing the second possibility, he may well leave. As a result, the producer will indeed try to keep the customer happy.

Long-term contracts that penalize the buyer when he changes sellers reduce the incidence of these departures. This mobility is also reduced if the buyer and seller integrate and the managers of the integrated enterprise mandate internal transactions. Both of these approaches make it more difficult to ensure high quality. The seller is now less worried about losing his customer, so it is harder to induce him to improve quality.² In addition, both approaches make it more likely that the buyer will continue to purchase from a seller whose competence has declined. That long-term contracts and integration are at least sometimes associated with low quality should be apparent to those familiar with the food services at most American universities.

The evidence of Robert Eccles and Harrison White (1988) also suggests that low quality and internal transactions go hand in hand. They interviewed managers who buy and sell goods across profit centers in firms with multiple profit centers. The impression left by these interviews is that managers prefer external to internal transactions.³ They would rather buy from and sell to agents outside the firm. Internal transactions occur only when they are mandated by management. Frictions are much more common in these mandated transactions than in their external counterparts. Eccles and White justly regard their findings as a challenge to Williamson's (1985) view of firms. Their evidence suggests that transactions costs actually increase when firms engage in internal transactions.

²According to Coase (1988), he discusses this loss in quality from integration in his 1934 notes for his paper "A Theory of Contract."

³Much as university professors prefer to eat in establishments that are not run by their employer.

If long-term contracts and integration make quality more difficult to deliver, why do they arise? The reason is that delivery of high quality in spot markets requires that price exceed marginal cost. Reductions in mobility thus raise the combined profits of buyer and seller. When the buyer leaves, he starts paying rents to others. These rents are captured by tying the buyer to the seller.⁴

The paper proceeds as follows. Section I presents a variant of the Klein and Leffler (1981) model in which the actual level of quality is stochastic even when the supplier is trying hard to accommodate the buyer. For simplicity, I focus here only on the randomness introduced by temporary bad luck on the part of the seller. Section II presents an illustrative long-term contract. The contracts described in this section are generally efficient. They both provide a benchmark for the inefficiencies studied in later sections and illustrate the rent-capture advantages of reducing buyer mobility.

Section III focuses on inefficient outright integration of the buyer and seller. Here, I suppose that there is an alternative mechanism that can also improve quality. This more costly mechanism works even if there is no threat of customer departure. For instance, costly managerial attention might be able to ensure high quality by making seller infractions verifiable by outside agents. Alternatively, costly managers might be given a great deal of discretion which, as in Alchian and Demsetz (1972), allows them to induce the seller to perform some quality-enhancing tasks. Even this may not raise quality as much as the effort brought forth by spot markets. Because this mechanism has higher cost per unit of quality, it is socially inefficient. Nonetheless, the wedge between price and marginal cost may make integration with these features privately profitable.⁵

⁴For other models in which socially inefficient contracts arise because they help capture rents, see Philippe Aghion and Patrick Bolton (1987), Mathias Dewatripont (1988), and Janusz Ordover et al. (1990).

⁵The wasteful monitoring by managers of their suppliers appears to be similar to the wasteful monitoring of workers in Samuel Bowles (1985). There, managers waste resources monitoring to avoid having to pay

Section IV introduces imperfect competition which raises prices. With spot markets alone, such increases in price have a socially desirable result. They make it more likely that producers will supply high quality, because they now have more to lose when the customers depart. Unfortunately, these high prices do not diminish the benefits from integration or the benefits from signing a long-term contract that eliminates all buyer departures and has low quality. As a result, increases in prices raise the likelihood that long-term contracts and integration are inefficient relative to the spot-market outcome.

Section V focuses on a different inefficiency of equilibrium long-term contracts. This inefficiency arises when the aptitude of the seller varies stochastically. Because the seller sometimes becomes inept, the immediate departures of customers when they encounter bad quality are socially beneficial. They raise the average quality that customers receive. Nonetheless, long-term contracts that reduce buyer mobility and raise the likelihood of purchases from inept suppliers can raise profits, because price exceeds marginal cost. Section VI presents my conclusions.

I. Spot Markets in the Klein-Leffler Model with Free Entry

Sellers produce a homogeneous intermediate good. The marginal cost of a bad-quality version of the good is c . After spending

workers an efficiency wage (which represents an excess of the wage over the disutility of labor). Our papers differ not only in focus but also in structure. In Bowles's model, agent A wastefully monitors agent B to extract rents from B. Assuming that contracts can specify the level of monitoring, the contract between A and B is suboptimal for A and B taken together. Overall efficiency would be restored if B were allowed to pay A *ab initio* to refrain from monitoring.

By contrast, in my model, A and B benefit jointly from writing a contract that involves monitoring. They both benefit because their arrangement extracts rents from C, a third party who is not involved in the contract. This difference is partly responsible for the models' different conclusions. Whereas in my model firms grow too big, they need never grow larger than their minimum efficient scale in Bowles.

ε , the quality of the good sometimes improves. I assume that a particular item becomes a high-quality good with probability σ if the firm spends an additional ε on the item. With probability $1 - \sigma$, the quality of the item remains poor.

There are Y customers who are, each period, willing to pay up to r for one unit of low quality. They are willing to pay an additional μ for a unit of high quality. Customers are risk-neutral: they are willing to pay $r + x\mu$ for a unit which is of high quality with probability x .

Contracts cannot specify the level of quality. This incompleteness of contracts has several possible sources. One possibility is that outsiders are unable to observe quality, so that contracts stipulating payments as a function of quality are unenforceable. Another possibility is that quality is difficult to spell out when contracts are written, although buyer and seller both recognize it when goods change hands.

Therefore, the qualities of the good I am concerned with here are not easily measurable. Rather, they are things like courteous service and prompt handling of complaints. Another important quality attribute in which these problems arise is seller flexibility in meeting changes in the buyers needs. Consider the following. The buyer and seller originally write a contract for delivery at the beginning of December. The buyer then realizes that he would much prefer delivery in the middle of November. A high-quality seller is flexible, so that he is capable of changing the delivery at a low cost. A seller of bad quality finds it very costly to change the delivery date. The courts see only the *ex post* cost of changes in delivery terms and not the effort made in being flexible.⁶

In this section, I consider spot-market transactions in which the only payments are proportional to the number of items exchanged. Each purchasing transaction has three stages. First, all upstream suppliers announce their prices. Second, buyers choose sellers and agree to receive one unit at this price. Finally, delivery occurs either

⁶This example of quality is very reminiscent of Grossman and Hart (1986).

with low or high quality. Customers know what quality they have received, though not the quality received by others.⁷

Customers' beliefs about the quality they can expect to receive depend on the price charged and on their personal experience. There are different equilibria depending on the precise form of these beliefs. The worst equilibrium for suppliers has customers believing that goods are universally of bad quality. As long as $r \geq c$, there is then an equilibrium where all active suppliers charge c and provide bad quality. For a producer to deviate by providing either good quality or lower prices represents a simple gift to customers. For him to deviate by raising his price loses him all sales.

Equilibria with high quality have different beliefs and higher prices. At these equilibria, customers believe instead that essentially all producers who charge a relatively high price P try to provide high-quality goods. In equilibrium, various firms charge P and are initially regarded as identical. Customers thus pick their initial supplier at random.

If this supplier delivers high quality and continues to charge P , they come back to this supplier in the following period. By contrast, if he delivers low quality, they choose a supplier at random from among the other producers who charge P . The assumption that customers leave suppliers from whom they have received low quality while they stay with those whose quality has been good is crucial for supporting equilibria where quality is high. As will become clear below, if customers returned no matter what quality they received, there would be no incentive to provide high quality.

The only formal justification offered in this section for the departure of unsatisfied customers is their indifference among all suppliers. In practice, it seems reasonable to suppose that customers expect suppliers

who have failed to provide good quality to be worse than the average supplier. As a result, customers want to remain only with those suppliers whose quality has been high. I model this formally in Section V.

I now study the incentives to provide high quality at equilibria of this kind. In these equilibria, the value to a firm from having a single customer is V . By contrast, the value to a firm of a customer who is currently buying from another supplier is W . Therefore the value V equals

$$(1) \quad V = \max\{P - c - \varepsilon + \delta[\sigma V + (1 - \sigma)W], P - c + \delta W\}$$

where δ is the discount factor. The first expression in (1) is the value when the firm tries to provide high quality, while the second expression is the value when it decides to provide low quality. If the number of firms is finite, the value to a firm of a customer who is currently buying from another firm is proportional to V . The reason is that this customer can be expected to buy from the current firm sometime in the future. Therefore, W equals ξV , where ξ is smaller than 1 since, at best, the customer comes back after one period. If customers who switch pick their firm at random, this parameter goes to 0 as the number of firms goes to infinity. Let V^h and V^l denote the values of making and failing to make the effort to improve quality. Then, (1) implies

$$V^h = \frac{P - c - \varepsilon}{1 - \delta[\sigma + \xi(1 - \sigma)]}$$

$$V^l = \frac{P - c}{1 - \delta\xi}$$

Attempting to provide high quality is worthwhile if V^h exceeds V^l or

$$(2) \quad \varepsilon \leq \delta\sigma \frac{1 - \xi}{1 - \delta\xi} (P - c).$$

It is apparent from this expression that it becomes easier to sustain high-quality equi-

⁷Klein and Leffler (1981) let customers who have received bad quality communicate this to other potential buyers. These then desert the producer as well. This is less appealing if customers differ in their experiences with the producer.

libria as P becomes higher. In particular, because δ , σ , and $(1-\xi)/(1-\delta\xi)$ are each smaller than 1, high-quality goods are sold only if

$$(3) \quad P - c > \varepsilon$$

so that the firms make positive profits. Klein and Leffler (1981) and their followers have addressed what they regarded as the uncomfortable coexistence of these rents with free entry. Carl Shapiro (1983), for instance, argues that initial expenditures are necessary to establish a reputation for good quality and that the rents compensate the firm *ex post*. However, the model is equally consistent with free entry in the absence of such initial expenditures.

Free entry does prevent the equilibrium price P from exceeding c by more than $\sigma\mu$:

$$(4) \quad P - c \leq \sigma\mu.$$

Otherwise, a firm would enter, undercut the price by $\sigma\mu$, and sell low-quality goods to Y customers.

If customers expect low quality from firms that charge less than P , entrants who charge between P and $P - \sigma\mu$ get no customers, and condition (4) is the only restriction on price. Suppose instead that customers believe that any firm charging a price such that (2) holds tries to provide high quality. Such beliefs seem reasonable; firms that charge such prices will indeed try to provide high quality if customer desertions occur only in response to low quality. In the presence of such beliefs, (2) cannot hold as a strict inequality. Customers gravitate to entrants whose price makes (2) hold as an equality.

Now consider the incentives of a firm with attached buyers to charge a price different from P . If it raises its price above P , all customers will leave, and it will lose its usual profits. Moreover, its own benefit from reducing its price below P is no bigger than a new entrant's profit from undercutting P . The reason is that such cuts in P reduce revenues from the firm's attached customers. The requirement that firms with attached customers must have no incentive

to deviate from the price P thus imposes no additional restrictions.

Combining (2) and (4), equilibria with high quality exist only if

$$(5) \quad \varepsilon \leq \delta\sigma^2\mu \frac{1-\xi}{1-\delta\xi}.$$

From a social point of view, the effort to provide high quality is worthwhile if and only if

$$(6) \quad \varepsilon < \sigma\mu.$$

Otherwise the benefit from the effort $\sigma\mu$ is smaller than the cost ε . Because δ , σ , and $(1-\xi)/(1-\delta\xi)$ are each smaller than 1, (5) implies (6). When it is not socially worthwhile to make the effort, it will not be provided at any equilibrium with free entry. On the other hand, (6) does not imply (5). Condition (6) holds, while (5) is violated when

$$(7) \quad 1 < \frac{\sigma\mu}{\varepsilon} < \left(\frac{1-\delta\xi}{1-\xi} \right) \left(\frac{1}{\delta\sigma} \right).$$

Then, the effort to improve quality is socially worthwhile but not provided at an equilibrium with free entry. The region described by (7) disappears as δ and σ go to 1 and ξ goes to 0. Underprovision of quality (relative to the social optimum) becomes more likely as customers leave more often for accidental reasons (low σ), take longer in coming back (low δ), or come back even after having been disappointed (high ξ).

II. An Illustrative Long-Term Contract

In this section, I illustrate the advantages and disadvantages of a particular form of long-term association whose main purpose is to make separations less likely. I suppose that a single buyer-seller pair can write a long-term contract. Whatever contract this particular buyer-seller pair writes, the relationship of all other buyers with their seller is unaffected. The spot market considered in the previous section with price P continues to operate.

A general long-term contract has at least three elements. First, there is an initial transfer. Then, there is a payment \hat{P} to be made from the buyer to the seller in each period that the buyer requests a unit from the seller. Finally, there is a severance payment s that the buyer makes to the seller when he changes suppliers. For simplicity, I assume that ξ is essentially zero; once this payment has been made, the buyer never returns to the seller.

The contracts in the previous section can be put in this form by letting the firm charge the same price as the others and letting the severance payment equal 0. Buyers were then indifferent between staying with their current supplier and changing suppliers. This indifference allowed me to assume that buyers always stay when they receive good quality and always depart when they receive low quality. Here, I assume instead that the contract also stipulates the probability λ that the buyer stays after he receives low quality. While there is no mechanism that enforces that buyers leave with this probability, I ensure that they are willing to do so by choosing s so that they are indifferent between staying and leaving.⁸

Suppose that the contract is devised so that the seller tries to produce high quality. In this case, all sellers produce the same quality, so the buyer must be indifferent between the payment streams associated with different sellers. In particular, the buyer involved in the long-term contract must be indifferent between severing the relationship immediately and severing the relationship only after remaining with the seller for one more period. This implies that

$$s + \frac{P}{1-\delta} = \hat{P} + \delta \left(s + \frac{P}{1-\delta} \right)$$

or

$$(8) \quad s = \frac{\hat{P} - P}{1-\delta}$$

⁸Since the aim of the long-term contract is to keep the buyer and seller together, the probability that the buyer stays after receiving good quality remains 1. This also makes it easier to induce the seller to make the effort necessary to improve quality.

so the present discounted value of payments is the same whether the buyer stays with this particular seller or pays the severance payments and then buys at the price P .

Then, the value to the seller from having a currently attached buyer is

$$V = \max \{ \hat{P} - c - \varepsilon + \delta [(\sigma + (1-\sigma)\lambda)V + (1-\sigma)(1-\lambda)s], \\ \hat{P} - c + \delta [\lambda V + (1-\lambda)s] \}$$

Therefore, the values of trying to provide high quality and of providing only low quality are

$$(9) \quad V^h = \frac{\hat{P} - c - \varepsilon + \delta(1-\sigma)(1-\lambda)s}{1 - \delta[\sigma + (1-\sigma)\lambda]}$$

$$(10) \quad V^l = \frac{\hat{P} - c + \delta(1-\lambda)s}{1 - \delta\lambda}$$

Trying to provide high quality is profitable only if V^h is at least as large as V^l or

$$(11) \quad \varepsilon \leq \delta\sigma \left(\frac{1-\lambda}{1-\delta\lambda} \right) [\hat{P} - c - (1-\delta)s]$$

Substituting (8) in (11), the producer then actually tries to produce high quality if

$$(12) \quad \varepsilon \leq \delta\sigma \left(\frac{1-\lambda}{1-\delta\lambda} \right) (\hat{P} - c)$$

which is independent of s and \hat{P} . A one-dollar increase in s raises the present discounted value of payments if the buyer leaves by one dollar. To preserve the buyers indifference, the present discounted value of payments if the buyer stays forever must increase by one dollar as well. Similarly, seller indifference requires that a one-dollar increase in s (which reduces the payoff from trying to provide high quality) be matched by a one-dollar increase in the present discounted value of receipts if the buyer stays on. Therefore, as long as (8) is satisfied, the precise values of s and \hat{P} do not affect the provision of quality.

Raising s while ensuring that \hat{P} satisfies (8) has the same effect as raising the initial transfer from buyer to seller. Both make the seller better off at the expense of the buyer. For simplicity, I thus consider contracts in which s is 0. The resulting initial transfer, whose value does not affect the subsequent analysis, depends on the bargaining strength of the two agents. If the seller can make a take-it-or-leave-it offer which if rejected leads to the sequence of spot contracts, the transfer is 0. If, instead, the buyer can solicit offers from competing sellers, the transfer is negative.

A problem does arise if buyers can sign long-term contracts that give them large initial transfers not just initially but also at later points. Suppose a buyer believes that next period he can sign a long-term contract from which he receives so much initially that the total present value of his payments equals $(c + \varepsilon)/(1 - \delta)$. From the buyer's perspective, this is equivalent to having access to a price of $c + \varepsilon$ from next period on; but, with P equal to $c + \varepsilon$, it is obvious from (12) that it is impossible to provide high quality in the current period. Buyer access to future contracts on good terms makes it impossible to induce the seller to provide high quality now. For equilibria with high quality to exist, buyers must believe that those who offer good terms in the future will provide only low quality.

The right-hand side of (12) falls when λ rises. Indeed, it is equal to 0 for λ equal to 1. This means that higher values of λ make it more difficult to induce the sellers to provide high quality. The highest value for λ that is consistent with attempts to provide high quality makes (12) hold with equality and equals

$$(13) \quad \frac{\delta\sigma(P - c) - \varepsilon}{\delta\sigma(P - c) - \delta\varepsilon}$$

A higher value for λ would require a greater incentive to provide quality in the form of a higher \hat{P} or a lower s . However, either, of these changes would break the buyer's indifference and would result in immediate departure by the buyer.

There is an important case in which the maximal λ given by (13) is 0. This is the

case in which customers believe that any firm whose price satisfies (2) tries to provide high quality. Then, in equilibrium, (2) holds with equality, so (13) equals 0. Greater values of λ are possible only if there is slack in condition (2). Only then are firms so eager to try to provide high quality that they continue to do so even if dissatisfied customers sometimes remain.

The next question is the value to the buyer and seller of varying λ . I compute this value assuming, as before, that the initial bargaining between buyer and seller is efficient. Therefore, only the effect of λ on the sum of buyer and seller surplus is relevant.

When the buyer purchases from the seller, the two together obtain surplus equal to $r + \sigma\mu - c - \varepsilon$. With probability $\sigma + (1 - \sigma)\lambda$, the buyer returns the following period, and surplus in that period is again $r + \sigma\mu - c - \varepsilon$. With probability $(1 - \sigma) \times (1 - \lambda)$, the buyer leaves. He then receives a present discounted value of benefits equal to $(r + \sigma\mu - P)/(1 - \delta)$ from then on. Combining these terms, the present discounted value of the benefits to buyer and seller equals

$$(14) \quad \frac{r + \sigma\mu}{1 - \delta} - \frac{c + \varepsilon + \frac{\delta(1 - \lambda)(1 - \sigma)}{1 - \delta}P}{1 - \delta[\sigma + (1 - \sigma)\lambda]} \\ = \frac{r + \sigma\mu - P}{1 - \delta} + \frac{P - c - \varepsilon}{1 - \delta[\sigma + (1 - \sigma)\lambda]}$$

The first term in the final expression is the surplus the buyer gets when he makes all his purchases in the spot markets. Thus, the second term captures the benefits from the association. In each period that the buyer remains, these benefits equal the difference between P and the cost of production $(c + \varepsilon)$. Inequality (4) says that this difference is positive. This implies that expression (14) is increasing in λ . This is the basic problem. The buyer and seller benefit

from increases in λ because such increases capture rents from the other sellers who, collectively, get $P - c - \varepsilon$ when the buyer leaves. On the other hand, if λ rises too much, there is nothing that keeps the seller from offering bad quality.

Using the maximal value of λ in (13), this present value becomes

$$(15) \quad \frac{r + \sigma\mu - P}{1 - \delta} + \frac{(P - c - \varepsilon)[\sigma(P - c) - \varepsilon]}{(1 - \delta)[\sigma(P - c) - \sigma\varepsilon]} = \frac{r + \sigma\mu - c - \varepsilon/\sigma}{1 - \delta}.$$

There does exist an empirically implausible method for letting λ , and thus joint profits, be bigger without sacrificing quality. This method, which I ignore here, is to have the buyer and seller write a contract with a third party. Such a third party could receive payments from the buyer (and even from the seller) when the buyer changes suppliers. Such payments tend to assure the seller of the buyer's fidelity without encouraging the seller to reduce his quality. In the current setup, reductions in quality are encouraged by the severance payments s since these go to the seller. The problem with schemes involving payments to third parties is the obvious incentive that they create for collusion between the third party and the seller.

The contracts considered so far are as efficient as the outcomes with spot markets, since they have the same quality. Within this class of contracts, inefficiency would be present only if the long-term contracts led sellers to forgo high quality. Then, joint profits when the buyer purchases from the seller equal $r - c$ per period. After the buyer leaves the seller, his net profits have a present discounted value of $(r + \sigma\mu - P)/(1 - \delta)$. Since the buyer leaves with probability $1 - \lambda$ in each period, the present

value of joint profits equals

$$(16) \quad r - c + \delta \left[\left(\frac{\lambda}{1 - \delta\lambda} \right) (r - c) + \left(\frac{1 - \lambda}{1 - \delta\lambda} \right) \left(\frac{r + \sigma\mu - P}{1 - \delta} \right) \right] = \frac{r + \sigma\mu - P}{1 - \delta} + \frac{P - c - \sigma\mu}{1 - \delta\lambda}.$$

Increases in λ are worthwhile only when the price P exceeds $c + \sigma\mu$. Then, the buyer would prefer low quality at a price of c to paying P and getting high quality with probability σ . However, as I showed in Section I, free entry implies (4) and is thus inconsistent with this case. Inefficient long-term contracts in which low quality prevails cannot be optimal responses to high-quality spot-market equilibria with free entry. They can only be optimal responses if, as in Section IV, (4) is violated and the equilibrium price exceeds its maximum free-entry level.

In this section, I have illustrated the tension between rent capture and high quality. I have also shown that long-term contracts yield efficient outcomes relative to spot-market contracts when there is free entry. In the next three sections, I explore three inefficiencies relative to the spot-market outcome. The first comes from the existence of an alternative technology for providing high quality, the second comes from prices in excess of the free-entry level, and the third comes from random variations in the competence of the sellers.

III. Outright Integration

Outright integration differs from the sort of long-term contract I considered in Section II in that it changes the control rights of various agents. Integration is probably most easily thought of as involving a new agent who has (limited) authority over both the buyer and the seller. Hiring such an agent is attractive if he can improve the quality that is forthcoming when the buyer

and seller let λ equal 1. The seller and buyer would then keep all the rents this buyer would ultimately have paid to other sellers.

Suppose that the hiring of this agent at a cost of m per period ensures that the seller makes the requisite effort when he produces for a particular buyer. Assuming this payment only occurs when the customer buys from the seller, the present discounted value of the joint benefits equals

$$(17) \quad \frac{r + \sigma\mu - m - c - \varepsilon}{1 - \delta\lambda} + \delta \left(\frac{1 - \lambda}{1 - \delta\lambda} \right) \left(\frac{r + \sigma\mu - P}{1 - \delta} \right) \\ = \frac{r + \sigma\mu - P}{1 - \delta} + \frac{P - c - \varepsilon - m}{1 - \delta\lambda}.$$

This expression is maximized for λ equal to one if

$$(18) \quad P \geq c + \varepsilon + m.$$

Otherwise, the optimal λ is 0, and integration is not valuable. Integration is worthwhile only if the cost of the internally produced good, $c + \varepsilon + m$, is competitive with the price of the externally supplied version.⁹ When (18) holds, the maximized value of (17) is

$$(19) \quad \frac{r + \sigma\mu - c - \varepsilon - m}{1 - \delta}.$$

This value exceeds (15), the value of the best long-term contract that keeps the cost of the quality effort equal to ε , if

$$(20) \quad m < \frac{1 - \sigma}{\sigma} \varepsilon.$$

⁹With free entry, (4) holds, so (18) implies that $\sigma\mu \geq \varepsilon + m$. More generally, unless this condition holds, the buyer and seller together are better off forgoing quality altogether, rather than paying $\varepsilon + m$ for a benefit of $\sigma\mu$.

Condition (20) becomes easier to satisfy as σ becomes smaller. The reason is that when σ is low there are many separations, so that the rents captured by buyer and seller jointly when their contract has λ given by (13) are relatively low. Then, their collective rents can be increased substantially by letting λ equal 1 and spending m per period. By contrast, when σ is 1 (the original case in Klein and Leffler [1981]), condition (20) is violated for all strictly positive m . In this case, buyer and seller never need to separate in equilibrium, so that nothing is gained by forcing them to remain together.

The combination of (20) and (2), which is necessary for spot markets to involve high quality, implies that $\delta(P - c)$ exceeds $\varepsilon + m$. This in turn implies (17). Thus, for any σ strictly less than 1, there exist strictly positive m 's such that (20) holds, and if the spot market equilibrium has high quality, integration is privately desirable and socially inefficient. If vertical integration could somehow be eliminated, quality would be the same, and society would save m .

Up to now, I have interpreted m as a managerial cost which ensures that the effort of the seller (whose cost remains ε) is actually forthcoming. I now discuss informally some other possibilities; more formal treatment is postponed for future work. One possibility is that spending m renders the seller's exertion of effort verifiable by outsiders. For instance, the manager might record the seller's actions in a way that outsiders trust. Then, a contract that stipulates that the seller be paid ε if he actually incurs the requisite effort becomes enforceable.

Along similar lines, the expenditure of m might produce verifiable evidence not on the effort expended but on the quality that the buyer actually receives. Then, the manager could write a contract with the seller in which the payments to the seller depend on the quality received by the buyer. Suppose the seller receives c when quality proves to be low and $(c + \varepsilon)/\sigma$ when it proves to be high. As long as he is risk-neutral, he is now happy to exert the necessary effort, since he on average earns $c + \varepsilon$ when he does and c when he does not.

The role of top management in the preceding two paragraphs is to spend m collecting evidence on quality (or on the effort to supply quality). Management is the monitor as in Alchian and Demsetz (1972).¹⁰ One can also think of m rather differently. The cost m can represent a reduction in quality itself, rather than a managerial cost in obtaining the spot-markets outcome. In this interpretation, the integrated firm produces a good worth $r + \sigma\mu - m$, rather than $r + \sigma\mu$. Even more generally, m can represent any loss in utility for buyer and seller (relative to $\sigma\mu - \varepsilon$).

One particular form of this loss stressed by Eccles and White (1988) is the prevalence of conflict in integrated organizations. They suggest that conflict between buyer and seller within an organization is useful, because it helps management gather information. I now suggest a scenario in which such conflict is indeed useful. Suppose that, just before the good is produced, the buyer receives a signal. This signal tells the buyer which particular change in specification would increase buyer utility and by how much. The buyer would now like the seller to make the change. On the other hand, the seller finds any such change costly. Moreover, the seller knows something about the cost of the change that others do not know. Therefore, the seller would normally overstate the costs of the change, either to increase his compensation or to avoid the necessary effort. In the absence of any mediator between buyer and seller, the result would be costly bargaining.

The manager can now serve a useful role. He can be put in charge of deciding whether the change should be made.¹¹ To find out whether the change is worthwhile, he encourages buyer and seller to "fight." In this fight, both buyer and seller present evidence. They also suggest experiments for

the manager to carry out. Since the manager has complete control over the tools with which buyer and seller work, he can actually carry out these experiments and learn something about the cost and value of the change. This knowledge then guides his decision. Note that the knowledge acquired by the manager is unlikely to be perfect. Therefore, it is in the seller's best interest to pretend that his costs are high to the very end.

The upshot of all this could well be a gain to the buyer g_1 which is lower than $\sigma\mu$ and a cost to the seller g_2 which exceeds ε . In this case, m is given by $\sigma\mu - \varepsilon + g_2 - g_1$. As long as (17) and (20) are met, setting λ equal to 1 and hiring the arbiter-manager is worthwhile for the buyer and seller taken together. Alternatively, an arbiter-manager will find it worthwhile to hire both a buyer and a seller. This is a more appealing description, because it emphasizes that, in the integrated organization, the buyer and the seller are employees of the arbiter-manager.

Suppose that these buying and selling employees carry out both internal and external transactions. Which will they prefer? When the buying employee carries out an external transaction, he receives $\sigma\mu$ in direct utility. This is the utility that results from courteous service and flexible specifications. Both of these facilitate the job of the buying employee. By contrast, when the buying employee carries out an internal transaction, he receives only g_1 in direct utility. Similarly, the selling employee's disutility from an internal transaction, g_2 , exceeds that from an external one. It thus appears to be possible that, consistent with Eccles and White's (1988) empirical findings, both buyer and seller prefer outside transactions.

Whether buying and selling employees are averse to internal transactions obviously depends also on how compensation varies with the type of transactions that they carry out. If employees are heavily compensated every time they buy or sell internally, internal transactions will look good to them. Suppose in particular that the buying and selling employees own the integrated firm and keep its profits. Since integration raises profits, net of the utility costs of buyer and

¹⁰Their model assumes there is a technological reason for monitors. As a result, the monitoring that emerges is socially good. By contrast, the monitoring that takes place here could usefully be replaced by a market.

¹¹This arbiter role of management is also stressed by Williamson (1975).

seller, the two would then, prefer internal transactions.

Suppose that, on the contrary, the two are only employees and the profits go to a third party. Then, at least on average, buying and selling employees must be compensated for the disutility induced by internal transactions. This implication is at least consistent with the evidence that wages rise with establishment and company size (see Charles Brown and James Medoff, 1989). In the view of this paper, employees of larger (more integrated) companies must be paid a compensating differential because they have to transact disproportionately more with other employees of the same organization.

This does not mean, however, that employees will be paid a bonus in every instance when they carry out an internal transaction. As long as agents are risk-neutral, there is little reason for compensation to vary in this manner. From an administrative point of view, such micromanaging of compensation is probably burdensome, since it requires precise definitions of the extent to which a transaction is internal. Thus, organizations may opt for flat salaries that reflect the *typical* volume of internal transactions. With this compensation structure, employees get more utility from external transactions than from internal ones. For internal transactions to take place at all, they must then be mandated by management.

One issue that arises at this point is why the cost m must be spent within an organization that also includes the buyer and the seller. Why can't these resources be spent in an arms-length transaction? The answer is that, in the situation I have in mind, arms-length transactions need very little management and supervision. The threat of dissolving the relationship is sufficient to ensure considerable effort by the seller.

Still, why can't the buyer and seller write a long-term contract that binds them to each other and then spend m on the services of an outside party? There are two reasons. The first, and principal, reason is that monitoring, policing, and verification require power. The monitor must be able to change the conditions of production to find

out how the seller and buyer are actually behaving. This flexibility is essential precisely because it is difficult to predict in advance the form of the effort to provide quality. Therefore, the monitor must be in a position to dictate how the assets with which buyer and seller work are used. Such residual control over the use of assets virtually defines the role of top management. Indeed, from a legal point of view, such residual control is the purview of the owner of the asset.¹²

Second, familiarity with the buyer and seller helps in determining the cost of various quality-enhancing actions, as well as in verifying that these actions have been taken. Obtaining this familiarity takes time. There is thus a benefit to having the monitor remain in his job for a long time. By making the monitor part of the firm, his mobility costs increase, and he becomes more likely to learn the personal characteristics of buyer and seller.

The advantages of mergers that I describe are similar to those of the "incomplete contracts" theory of the firm surveyed by Holmstrom and Tirole (1987). As in that article, some individuals are put in charge, because contracts cannot specify all contingencies. These individuals get to decide how to proceed when the contract does not specify what is to follow.

Grossman and Hart (1986) also base their theory of the firm on contractual incompleteness. Yet, the implications of their model appear to be rather different from the one presented here. They focus on the contractual incompleteness that makes it difficult for individuals to garner the fruits of their human capital accumulation. In the absence of ownership rights, some of the benefits of an individual's human capital accumulation accrue to those he works with. Ownership can ameliorate this. The owner gets to decide on all aspects of the transaction not specifically contracted on in advance. As a result, he receives a disproportionate share of the payoff to both his and

¹²Grossman and Hart (1986) equate residual control over the use of assets with ownership.

his underlings human capital accumulation. Ownership of an asset by agent A alleviates the moral-hazard problem inherent in A's human capital accumulation (while worsening that of the other agents who work with this asset). They are careful to present their model as one of an owner-manager, though their theory perhaps applies also to corporate top management. In this case, one would expect top managers' compensation to be closely related to firm performance.

In my story, by contrast, moral hazard may not be an issue for top managers. Top managers are like policemen. The competence with which they dispatch these functions may be easy to determine from their reports. Top managers could still earn large sums if the skills needed to investigate whether the appropriate effort is being carried out are in short supply. In this case, one would expect larger firms, whose management problems are more complex, to hire more-competent monitors. This suggests that management compensation should increase with firm size. By contrast, their wages need not bear any particular relationship to firm performance. In practice, managerial compensation does seem more linked to firm size than to firm performance (see Michael Jensen and Kevin Murphy, 1988).

One critical question faced by all theories of the firm is the conundrum posed by Coase (1937) and Williamson (1985): what limits firm size? If top managers can intervene selectively and obtain whatever advantages there are to integration, a single firm should control the whole economy. This argument gains additional force when one recognizes that integration eases collusion in product markets.

Williamson (1985) gives a variety of reasons why managers cannot actually intervene so selectively. One reason that appears particularly germane in my context is that managers are imperfect and sometimes make mistakes. For managers to be effective at policing, they must have great authority, so these mistakes can be costly. They become more costly the larger is the manager's span of control. Giving a manager residual decision rights over the assets used by many individuals when his decisions only rarely

contribute to the common good is dangerous. By contrast, giving him these rights when intervention is needed often is more attractive.

IV. Imperfect Competition

Instead of letting entry be free, I assume here that the number of sellers is fixed at N . I show that the extent to which integration is inefficient can then depend on the level of rents in the industry.

Without free entry, price can exceed $c + \sigma\mu$. It can also exceed the level that makes (2) hold with equality even when customers view all firms whose price satisfies (2) as trying to provide good quality. To sustain such high prices, firms threaten each other with punishments as in James Friedman (1971). The worse the threatened punishment, the higher the prices that can be sustained in equilibrium. For concreteness, I focus on the worst possible punishments in the style of Dilip Abreu (1986). Similar results obviously obtain for weaker punishments.

The structure of the equilibrium is the following. Each firm is expected to charge P and earn a present discounted value of $\pi = Y(P - c - \varepsilon)/N(1 - \delta)$. A firm that deviates by raising its price loses all its customers, so this deviation is unattractive.

Suppose a firm deviates by cutting its price. Then, a punishment period ensues. All other firms charge $c + \varepsilon$ and try to provide high quality in the period after the deviation. In this period, the deviating firm is expected to charge a price L so low that its losses in that period equal $\delta\pi$. If firms charge these prices, they revert to charging P in the subsequent period. If any firm charges a different price, it is itself punished in analogous fashion.

Several comments about this construction are in order. First, all firms have an incentive to provide high quality if satisfied customers return in the following period, when the price is again equal to P . Second, the firm that is being punished can expect to earn a present value of 0 whether it charges L or not. If it does charge L its present discounted value of profits is 0 by construc-

tion. If it deviates and charges a higher price, customers expect its quality to be low, so that they are willing to pay at most c for its goods. This means that the deviating firm can do no better than charge c , break even in that period, and get punished again. This again gives it 0 in present value.

By cutting its price slightly below P , a firm can, at most, capture the entire market. Then, it would earn N times its usual per-period profits in the period of the deviation. After that, it earns a present value of 0. Thus, these deviations will not take place as long as

$$N \leq \frac{\delta}{1 - \delta}$$

which is independent of P and is satisfied for sufficiently high δ .

As in the case of free entry, there are both equilibria where quality is low and equilibria where quality is high. However, if (6) is met, switching from an equilibrium with low quality to one in which the firms try to supply high quality increases willingness to pay by more than it increases costs. Moreover, such a switch does not increase the equilibrium's vulnerability to deviations. Therefore, if the selling firms can pick equilibria and (6) is met, they will pick the equilibrium in which they produce high quality and charge $r + \sigma\mu$.

This ability of sellers to coordinate on equilibria with high prices is socially beneficial. It means that condition (2) becomes easier to meet, and firms find it in their interest to provide high quality. In particular, (2) can now be met even when (5) fails (so that free-entry equilibria have low quality). Moreover, in this model, there is no other distortion from high prices. With spot markets alone, their only effect on economic efficiency is to make high-quality equilibria possible in situations where, with free entry, only low-quality equilibria exist.¹³

¹³Contrast this with Michael Spence (1975), where monopoly (and thus collusion) generally distort quality away from the optimum.

I now study the effect of high prices on the attractiveness for one buyer-seller pair of both long-term contracts and outright integration. Consider first long-term contracts of the form considered in Section II. It is apparent from (14) that the derivative of joint profits with respect to λ rises with P . Higher prices make reductions in separations more attractive. As shown in (13), they also make higher values of λ feasible.

On the other hand, increases in P have a direct negative effect on the buyer. The net effect of these contradictory forces is that the joint profits at the optimal λ given by (15) are independent of P . The same is true of (16) with λ equal to 1, which gives joint profits if buyers and sellers forgo high quality altogether, and of (19), which gives joint profits under integration.

Therefore, the level of P itself has no effect on the choice of contractual form. However, it does affect the social optimality of the chosen contract. Suppose that ε is between $\sigma\mu$ and $\sigma^2\mu$ so that (5) is violated and market transactions in the presence of free entry involve low quality. If m is smaller than $\sigma\mu - \varepsilon$, buyer and seller will integrate. Relative to the spot-market outcome with free entry, this integration is efficient.

When ε is in this range, (6) is satisfied, so attempting to provide high quality is efficient. Moreover, if sellers with market power can coordinate, they choose a spot-market equilibrium with high prices in which quality is high; (16) with λ equal to 1 exceeds (15). Therefore, the contract with low quality, which avoids all separations, has higher joint profits than either spot-market transactions or any other contract of the form studied in Section II. If m is smaller than $\sigma\mu - \varepsilon$, joint profits are even higher under integration. However, both of these arrangements are socially less desirable than the spot-market equilibrium with high prices. Either they involve inefficiently low quality or inefficient integration.

The intuition for the result that the presence of market power reduces the efficiency of private contractual arrangements is the following. Market power raises prices, which makes the provision of high quality in spot markets easier. However, market power

does not reduce the private desirability of contractual forms where high quality, if it is provided at all, is costly.

V. Inefficient Separations with Long-Term Contracts

The long-term contracts of Section II feature fewer separations than the spot markets of Section I only because I have defined spot markets to have customers leave every time they encounter low quality. This definition of spot markets is, up to this point, arbitrary. If all producers are identical, customers ought to be indifferent among them. Thus, it is just as consistent with consumer maximization to assume that customers stay with a probability given by (13) when they have received low quality. This still ensures high quality and is identical to the long-term contracts of Section II.¹⁴ Another consequence of this indifference is that, at least with free entry, long-term contracts are just as socially efficient as what I have dubbed spot-market transactions.

In this section, I provide a reason for customers to leave when they encounter low quality. The reason is that low quality can mean two things. Either the seller has been temporarily unlucky (as I have assumed until now) or the low quality is an indication of seller incompetence.¹⁵ This possibility of seller incompetence makes buyers want to switch sellers whenever quality is low.

This possibility implies that long-term contracts that attach buyers to sellers are subject to a new form of inefficiency. This inefficiency is present even when sellers always try to provide high quality. It stems from the increased probability that buyers will purchase from incompetent sellers. In

spite of this inefficiency, such attachments can be privately optimal when price is high.

I suppose that firms can be in one of two states. In state w , they are well managed while in state b they are poorly managed. Poorly managed firms are intrinsically unable to provide high quality. They spend ε , and their quality is low nonetheless.¹⁶ The probability that a well managed firm remains well managed in the next period is $1 - \phi$. The probability that a poorly managed firm becomes well managed is $1 - \psi$. I assume that ψ exceeds ϕ , so that bad management has a tendency to persist. These transition probabilities imply a steady-state probability of being in state w equal to

$$(21) \quad \eta = \frac{1 - \psi}{1 + \phi - \psi}.$$

For simplicity, I also let the initial probability of being in state w equal η . Because $\psi > \phi$, η is smaller than the probability $1 - \phi$ that a previously well managed firm remains well managed.

Suppose that there is a spot-market equilibrium in which all firms charge P and in which well managed firms try to supply good quality. I first consider customer behavior, assuming such an equilibrium exists. I prove that customers come back to sellers who have given them high quality, while they desert those that have not. I then give conditions for this equilibrium to exist. These conditions ensure that well managed sellers do indeed try to provide high quality and that no seller wants to charge a different price.

Consider a customer who receives high quality in period t . He infers correctly that his current seller is well managed. His probability of obtaining high quality next period from this particular seller is thus $(1 - \phi)\sigma$. By contrast, his probability of obtaining high quality from a seller chosen at random is $\eta\sigma$, which is smaller. He therefore prefers to remain with his old seller.

¹⁴Of course, if customers stay with higher probability, equilibrium quality must be low.

¹⁵The presence of two analogous possibilities also induces customer departures in the monopoly model of Allen and Faulhaber (1988). Because they focus on the case with a single supplier, the possibility of seller incompetence plays a more crucial role. Without it, customers do not react at all to low-quality realizations, so that high quality is not sustainable in pure-strategy equilibria.

¹⁶Letting them avoid the cost ε changes the formulas below without affecting the qualitative results.

Suppose instead that the customer receives low quality at t . The posterior probability that he assigns to having a badly managed seller depends on his prior probability. This prior probability is lowest if the customer has bought a high-quality good from this particular seller at $t-1$, so that he knows the seller was well managed at $t-1$. I consider this case first.

Let w_t represent the event that the seller is well managed at t , b_t the event that he is poorly managed at t , and L_t the event that he provides low quality at t . Then, the probability of event w_t given that the seller was previously well managed and that he provides bad quality at t is

$$\begin{aligned}
 (22) \quad P(w_t | L_t, w_{t-1}) &= \frac{P(w_t, L_t | w_{t-1})}{P(b_t, L_t | w_{t-1}) + P(w_t, L_t | w_{t-1})} \\
 &= \frac{(1-\sigma)(1-\phi)}{\phi + (1-\sigma)(1-\phi)}.
 \end{aligned}$$

This probability is smaller than η as long as

$$(23) \quad \sigma \geq \frac{\psi - \phi}{1 - \phi}.$$

This condition is met as long as σ is sufficiently large. Then, well managed firms are sufficiently likely to deliver high quality that low quality is a signal of poor future prospects.

Now consider a buyer who has had no previous experience with this seller because he has just changed suppliers. His prior probability that the seller is well managed is then η . The posterior probability that he is well managed given that he has provided bad quality is

$$\begin{aligned}
 (24) \quad P(w_t | L_t) &= \frac{P(w_t, L_t)}{P(b_t, L_t) + P(w_t, L_t)} \\
 &= \frac{(1-\sigma)(1-\psi)}{\phi + (1-\sigma)(1-\psi)}.
 \end{aligned}$$

When ψ exceeds ϕ , this probability is smaller than η , so that customers who have just arrived leave suppliers who give them bad quality. In conclusion, when ψ exceeds ϕ and (23) is met, all buyers leave firms that offer low quality.

I now turn to seller behavior. I show that, given that consumers leave those who provide bad quality, sellers find it in their best interest to supply good quality. I later turn to sellers' incentives to change prices.

For simplicity, I assume again that ξ is zero, so that customers who leave essentially never return. Then, the value to a badly managed firm of an attached customer, who is about to leave, equals $P - c - \varepsilon$. Therefore, the value V to a well managed firm of having a customer is

$$\begin{aligned}
 V &= \max\{P - c - \varepsilon + \delta[\sigma(1-\phi)V \\
 &\quad + \phi(P - c - \varepsilon)], P - c\}.
 \end{aligned}$$

The values of trying to provide high quality and of providing only low quality are now

$$\begin{aligned}
 V^h &= \frac{(1+\delta\phi)(P - c - \varepsilon)}{1 - \delta\sigma(1-\phi)} \\
 V^l &= P - c
 \end{aligned}$$

so that trying to provide high quality is profitable if

$$(25) \quad \varepsilon \leq \frac{\delta[\sigma + \phi(1-\sigma)]}{1 + \delta\phi}(P - c).$$

As long as this modified version of (2) holds, well managed firms do not deviate by cutting quality.

For such prices to be equilibria with free entry, customers must prefer to pay P and receive high quality with probability $\eta\sigma$, rather than to pay c and receive low quality for sure. Thus,

$$(26) \quad P - c \leq \eta\sigma\mu.$$

For equilibria with high quality to exist when entry is free, prices must exist that satisfy both (25) and (26). In other words,

$\eta\sigma\mu$ must exceed

$$(1 + \delta\phi)\varepsilon / \delta[\sigma + \delta(1 - \sigma)].$$

If, in addition, customers believe that firms whose price satisfies (25) try to provide high quality, the unique equilibrium has (25) holding with equality.

The incentives of firms with attached customers to raise price above P are more complex in this case. The reason is that customers who have received high quality are willing to pay more to their current firm than the price charged by other firms. To simplify the exposition, I eliminate this scope. I do this by assuming that customers believe that any firm that charges a price different from P will produce low quality. Raising price then leads all customers to depart.¹⁷

Note that the analysis of Sections I, II, and III corresponds to the model of this section when one takes the limit as ϕ goes to zero. At this limit, the firm is almost always well managed, but if σ exceeds ψ , (23) is met, and the buyer leaves any firm whose quality proves low. Moreover, at this limit, the payoff from the long-term contracts described in Section II is well approximated by (14).

Suppose, however, that the system is not at this limit and that a certain buyer-seller pair is given the option of signing a long-term contract. Privately optimal long-term contracts are now more complex. The reason is that information about a seller's competence accumulates over time. A seller who has delivered bad quality for two periods in a row is more likely to be poorly managed than one who has delivered bad quality for only one period. These complications preclude the derivation of the privately optimal long-term contract in this paper. Rather, I focus on a very simple type of long-term contract, and I show that it will be attractive

to buyer and seller taken together even though it is socially harmful.

This simple contract applies only to the first period. If the buyer receives high quality in the first period, he stays as before. If, on the other hand, he receives low quality, he now remains with probability λ . Whatever the reason for staying, his behavior in subsequent periods is the same as his behavior with spot markets; he leaves any seller whose quality is low. Thus, the contract gives the initial seller only one "second chance." If he delivers low quality after that, he loses the buyer.

To focus on the new inefficiency, I ignore contracts in which sellers cease to make the effort necessary to improve quality. In other words, I consider contracts whose λ is sufficiently small that the seller continues to try to provide high quality. By continuity, positive λ 's that satisfy this condition exist when (25) and (26) are satisfied. This leaves open for further research the question of whether high quality is easier to sustain for a given λ when, as in this section, ϕ exceeds 0.

From the perspective of buyer and seller together, this contract has a cost and a benefit. The cost is that, if the seller is indeed incompetent in the first period, it is costly for the buyer to stay. The benefit is that, if the seller is only unlucky in the first period, the contract ensures the buyer-seller pair some surplus that would otherwise go to some other firm.

The contract only matters when the buyer receives low quality in the first period, as occurs with probability $1 - \eta\sigma$. In the absence of any contract, the buyer would then leave and obtain good quality with probability $\eta\sigma$ from his new supplier. His total payoff after the initial period would equal

$$(27) \quad \eta\sigma(r + \mu - P + \delta G) \\ + (1 - \eta\sigma)(r - P + \delta B)$$

where G and B represent the present discounted value of the buyer's payoff after having received good and bad quality respectively.

If the buyer stays after receiving low quality in the first period, he has a probability ν

¹⁷This feature could be avoided by focusing on equilibria in which new customers pay lower prices (introductory offers) than do customers that remain. The basic result is not sensitive to this change in specification.

of facing a well managed seller where:

$$\nu \equiv P(w_l|L_l)(1-\phi) + [1-P(w_l|L_l)](1-\psi).$$

This expression is smaller than η , because the firm that has delivered low quality is more likely to be badly managed than a firm chosen at random. If the buyer who remains pays P , his total payoff from remaining is

$$(28) \quad \nu\sigma(r + \mu - P + \delta G) \\ + (1-\nu\sigma)(r - P + \delta X).$$

Equation (28) is identical to (27), except that the probability of the good outcome is $\nu\sigma$ instead of $\eta\sigma$. For λ to be positive, the buyer must be indifferent between leaving and staying. Thus, if he leaves, the severance payment must equal the difference between (27) and (28).¹⁸

Ignoring the initial transfer, the buyer's total loss from the contract is the difference between (27) and (28) times the probability $\lambda(1-\eta\sigma)$:

$$(29) \quad \lambda(1-\eta\sigma)(\eta-\nu)\sigma[\mu + \delta(G-B)].$$

A tedious but straightforward calculation leads to a value for $G-B$:

$$G-B = \frac{\delta(1-\phi-\eta)\sigma\mu}{1-\delta\sigma(1-\phi-\eta)}.$$

Since $1-\phi-\eta$ lies between 0 and 1, this expression is positive and so is the value of (29). Note that the loss (29) is independent of P and that it vanishes as ϕ goes to 0 because, in this case, both η and ν go to 1. While I computed (29) as the loss the buyer

from the lower quality he receives on average, it also equals the social loss from the contract. The reason is that the total social costs of production are the same whether the buyer stays or leaves, so the change in consumer surplus (at unchanged prices) equals the change in total surplus.

The private gain from the contract is that the price P exceeds the cost of producing these goods. After giving the buyer bad quality, there is still a positive probability ν that the seller is well managed. If the buyer stays, the buyer-seller pair therefore gains $P-c-\varepsilon+\delta V^h$ with probability ν . With 1 minus this probability, they gain only $P-c-\varepsilon$. Total gains from the excess of price over cost are thus given by

$$(30) \quad \lambda(1-\eta\sigma)(P-c-\varepsilon+\nu\delta V^h).$$

This expression is positive and increasing in P . For P sufficiently high or ϕ sufficiently low, (30) exceeds (29) for strictly positive λ . Then, it is worthwhile to make λ positive, thus incurring the social cost (29) in exchange for the rents (30). While this only demonstrates the private advantages and social costs of a very special long-term contract, similar costs and benefits attend more general long-term contracts.

VI. Conclusion

Ever since Coase (1937), economists have generally viewed the internal organizations of capitalistic firms as benign. Even Williamson (1975, 1985), who feels that vertical integration sometimes reduces competition, asserts that the main purpose of integration and long-term contracts is to reduce "transactions costs."

This paper suggests a less sanguine view. Whether the organization of firms is socially helpful is an empirical question. It is not a question economists should feel capable of answering on a priori grounds alone. Since little other information on the actual functioning of these institutions exists, economists should pay close attention to the comments of managers and to their dislike of internal transactions.

¹⁸This computation assumes that the price that the buyer pays if he stays is the same as the external price. Increases in the price that the buyer pays if he stays must be matched one-for-one with increases in the severance payment to maintain buyer indifference. Increases in price must also be matched one-for-one to keep constant the seller's incentive to provide high quality. Thus, once again, the breakdown between the price the buyer must pay and the severance payment does not matter.

There may, of course, exist other reasons for this aversion. A partial explanation is that there exist specialized individuals (or organizations) with great skill. These individuals are efficiently utilized only when they serve a great many different customers, so that their integration with customers is worthwhile only in the case of very large customers. A relatively small firm then has a choice of integrating with a less competent individual (to the dismay of its buyers) or having an external transaction with one of the specialized individuals. This story explains the preference of buyers for outside partners but does not explain the analogous aversion of in-house sellers.

Perhaps managers dislike intrafirm transactions because they cloud managerial compensation. Insofar as transfer prices are somewhat arbitrary, wages that depend on profits reported by profit centers become risky as intrafirm transactions become more common. Holmstrom and Tirole (1987) explain GM managers' dislike of centralized procurement along analogous lines. This story is not complete. Firms must have some reason for exposing their employees to additional risk. In the case of GM, it seems plausible that centralized procurement allows GM to use its monopsony power more effectively. Rent capture may thus lie behind GM's move.

My model has several potentially testable implications. First, in many industries, firms that provide their own inputs coexist with others that obtain theirs in the open market. These differences must involve some heterogeneity; the cost m of carrying out transactions internally must vary across firms. Insofar as these differences in m are exogenous differences in transactions costs, my theory has the same implication as Coase (1937). Suppose instead that m is the cost from a given degree of quality deterioration. Then, my theory predicts that buyers who are more concerned with quality ought to be particularly averse to integration. This suggests that, controlling for all other observable characteristics, the producers of "high-end" products within an industry ought to be less integrated than the producers of "average" products.

Second, according to my theory, downward vertical integration should be more prevalent in industries in which price exceeds marginal cost by more. Along more standard lines, the increased incentive to integrate when prices are high may be attributed to the desire to avoid the "double marginalization" problem. However, unlike my theory, this standard motivation for vertical integration loses its force if agents are able to write complex contracts such as two-part tariffs.

The theory presented here may also be helpful in understanding why the organization of transactions differs across nations. Perhaps these organizational differences are due to cultural differences. In my model, buyer mobility is less important in nations where people find it culturally less acceptable to provide low quality. Therefore, my theory predicts that long-term attachments are prevalent where these transactions costs are low.

Another possibility is that these organizational differences can be traced to other differences in the production process. Perhaps, the Japanese *kanban* system inherently leads to higher quality because of its use of small batches. If this is interpreted as a high value for σ , my model would predict relatively little centralized control of the whole production process and relatively extensive attachment via long-term contracts. While this suggests some of the features of *keiretsu*, extensive investigation of the extent to which the model can explain international differences is left for further research.

REFERENCES

- Abreu, Dilip, "Extremal Equilibria of Oligopolistic Supergames," *Journal of Economic Theory*, June 1986, 39, 191-225.
- Aghion, Philippe and Bolton, Patrick, "Contracts as a Barrier To Entry," *American Economic Review*, June 1987, 77, 388-401.
- Alchian, Armen and Demsetz, Harold, "Production, Information Costs and Economic Organization," *American Economic Review*, December 1972, 62 777-95.
- Allen, Franklin and Faulhaber, Gerald, "Opti-

- mism Invites Deception," *Quarterly Journal of Economics*, May 1988, 103, 397-407.
- Bowles, Samuel**, "The Production Process in a Competitive Economy: Walrasian, Neo-Hobbesian and Marxian Models," *American Economic Review*, March 1985, 75, 16-36.
- Brown Charles and Medoff, James L.**, "The Employer Size Wage Effect," *Journal of Political Economy*, October 1989, 97, 1027-59.
- Coase, Ronald H.**, "The Nature of the Firm," *Economica*, November 1937, 4, 386-405. (Reprinted in G. J. Stigler and K. E. Boulding, eds., *Readings in Price Theory*, Homewood, IL: Irwin, 1952.)
- _____, "The Nature of the Firm: Meaning," *Journal of Law, Economics and Organization*, Spring 1988, 4, 19-33.
- Dewatripont, Mathias**, "Commitment through Renegotiation-Proof Contracts with Third Parties," *Review of Economic Studies*, July 1988, 55, 377-91.
- Eccles, Robert G. and White, Harrison C.**, "Price and Authority in Inter-Profit Center Transactions," *American Journal of Sociology*, 1988, 94 (suppl.), S17-S51.
- Friedman, James W.**, "A Non-Cooperative Equilibrium for Supergames," *Review of Economic Studies*, January 1971, 38, 1-12.
- Grossman, J. Sanford and Hart, Oliver D.**, "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy*, August 1986, 94, 691-719.
- Holmstrom, Bengt R. and Tirole, Jean**, "The Theory of the Firm," in R. Schmalensee and R. Willig, eds., *Handbook of Industrial Organization*, New York: North Holland, 1989, 61-133.
- Jensen, Michael C. and Murphy, Kevin J.**, "Are Executive Compensation Contracts Structured Properly?" Harvard Business School Working Paper, February 1988.
- Klein, Benjamin and Leffler, Keith B.**, "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*, August 1981, 89, 615-41.
- Ordover, Janusz A., Saloner, Garth and Salop, Steven C.**, "Equilibrium Vertical Foreclosure," *American Economic Review*, March 1990, 80, 127-42.
- Shapiro, Carl**, "Premiums for High Quality as Returns to Reputations," *Quarterly Journal of Economics*, November 1983, 98, 659-80.
- Spence, Michael**, "Monopoly, Quality and Regulation," *Bell Journal of Economics*, Autumn 1975, 6, 417-29.
- Williamson, Oliver E.**, *Markets and Hierarchies*, New York: Macmillan, 1975.
- _____, *The Economic Institutions of Capitalism*, New York: Macmillan, 1985.

Advertising in Competitive Markets

By MARK STEGEMAN *

In this paper, small firms sell a homogeneous good to small consumers under conditions of free entry, but consumers receive price information only through firms' advertising. In equilibrium, every firm on the continuous price distribution buys less advertising than is socially optimal. The result is robust if firms advertise in just one medium. If readers of different advertising media are positively correlated, excess advertising can occur in media used exclusively to advertise discount prices. (JEL 024, 026).

Critics portray advertising as a socially pointless attempt to capture market share through psychological manipulation. Others defend advertising by emphasizing its informative role. Identifying the purpose of advertising does not, however, resolve the underlying welfare question. Even purely informative advertising can be purchased to excess. The issues are complicated enough to make welfare results scarce, but George Stigler and Gary Becker (1977) and Len Nichols (1985) (henceforth, SBN) produce a striking result. They show that firms in competitive markets buy the socially optimal quantity of advertising.¹

The SBN result is important but less general than it first appears. In particular, SBN's advertising is not informative in the usual sense: every agent in the SBN model knows exactly what commodities and attributes are available, where, and at what prices, regardless of the level of advertising.

In other words, SBN's advertising increases demand not by resolving uncertainty or improving access to markets but, rather, by improving the attributes of already traded commodities. Such advertising is essentially *persuasive*.²

This paper reexamines the welfare question for purely *informative* advertising. The markets investigated are competitive in the sense that small firms sell a homogeneous product to small consumers, but all information about price and product availability flows through advertising. Advertising has no other purpose. In this setting, every firm advertises less than is socially optimal. The result is quite robust. If ads are homogeneous *ex ante* (i.e., ad campaigns of comparable size reach similar consumers, on average), then every firm underadvertises, regardless of the composition of the agents in the market and the advertising technology.

Certain circumstances can overturn the underadvertising result, but they appear to be rather limited in nature. Section III describes one possibility: if ads appear in diverse media and the readerships of these media are positively correlated, then firms

*Department of Economics, University of North Carolina, Chapel Hill, NC. I am grateful to Peter Diamond, James Friedman, Eric Maskin, Torsten Schmidt, and two anonymous referees for helpful comments. The comments of one referee improved the formulation of Proposition 1. The usual disclaimer applies.

¹Nicholas Kaldor (1950) and Peter Steiner (1966) present arguments against advertising. Lester Telser (1964) and Dean Worcester (1976) present arguments in its favor. Howard Beales et al. (1981) summarize the issues. Stigler and Becker (1977) emphasize the attribute model itself; Nichols (1985) emphasizes its welfare implications.

²Of course, it is possible that SBN's advertising improves commodity attributes by improving consumers' information. For instance, advertising may improve aspirin's attributes by telling consumers about new uses for aspirin. Nevertheless, this seems to be logically indistinguishable from what is normally called persuasive advertising. In either case, consumers get higher consumption value for an existing commodity; who can judge whether that value is "real" or "imaginary"?

may advertise too much in media used exclusively to advertise discount prices. Another possibility arises from consumer search. In this paper, consumers get all price information through advertising. Elsewhere I show that, if consumers also search for price information, then firms advertise too much if consumer search costs are sufficiently small (Stegeman, 1986, 1990). Despite these qualifications, the underadvertising result is robust enough to be interesting: competitive firms always underadvertise if all price information travels through a single advertising medium, and they tend to underadvertise in other cases.

The competitive model should serve as a starting point for the analysis of noncompetitive markets: markets having large firms, heterogeneous products, or barriers to entry. For instance, Stahl (1989) shows that large firms underadvertise in a symmetric mixed strategy equilibrium,³ and in Stegeman (1986) I show that small firms generically underadvertise in the standard circle model of product differentiation (e.g., Stephen Salop, 1979). These results are discussed in Section IV.

To put the present work in context, it is instructive to review the SBN model in greater detail. Following Lester Telser (1964), SBN assume that consumers obtain utility from attributes they consume. Advertising is an input that enhances the attributes of the final product and is thus an implicit argument in the utility function. If the number of firms in the economy is large relative to the number of attributes, then the markets for attributes may be effectively competitive, and the standard welfare result applies. Competition implies the efficient

use of all inputs, including advertising. SBN emphasize that the result requires competition in the implicit markets for attributes but not in the markets for commodities.⁴

SBN's optimal-advertising result rests on four strong assumptions. The first assumption, that attribute markets are effectively competitive, appears to be at least as strong as the assumption that commodity markets are competitive, since the world appears to have at least as many attributes as commodities. The second assumption is that advertising has no effects on persons who do not buy the product. Third, SBN intentionally exclude the possibility that advertising persuades consumers to buy goods that they should not buy. Not all economists agree that such irrational or compulsive behavior should be excluded from consideration. Fourth, SBN's advertising conveys no information about the terms of trade, because everyone knows the implicit price of every attribute regardless of the level of advertising.

This paper differs in the first and fourth assumptions. The difference in the first assumption is not crucial. Although this paper emphasizes competition in commodity markets and SBN emphasize competition in attribute markets, either model is consistent with competition in both commodity and attribute markets. The important difference lies in the fourth assumption.

In this paper, the only purpose of advertising is to allow each firm to announce its existence and price. The fraction of consumers who see at least one of the firm's ads is a continuous, strictly increasing, and strictly concave function of the quantity of ads sent. If the unit cost of advertising is constant, then it follows that the total cost of advertising is a convex function of the number of consumers reached.

Gerard Butters (1977) studies such an advertising technology and shows that its use implies the failure of the law of one price. The argument is that a firm advertising a price at a jump in the price distribu-

³Stahl's (1989) model is similar to the simplest (i.e., Section I) version of the present model, except that his firms enjoy economies of scale in advertising and his consumers have downward-sloping demand curves. His results apply only to symmetric mixed-strategy equilibria, unlike the asymmetric equilibria of the present model. Stahl also investigates the impact of changes in the numbers of firms and consumers. His paper and the present paper were developed independently. Most of the results in this paper appeared originally in Stegeman (1986).

⁴Richard Schmalensee (1972) and others have made somewhat similar arguments.

tion could obtain a nonnegligible increase in sales through an arbitrarily small price cut; therefore, the equilibrium price distribution must be continuous. The argument is quite general. The equilibrium price distributions in this paper are continuous for the same reason.

Returning to the welfare question, there is no obvious reason to expect Butters's firms to provide the socially optimal quantity of informative advertising. Butters shows that firms do advertise optimally in his model, but his assumptions are very restrictive, and he provides no intuitive explanation. Others have recognized the presence of opposing externalities. One externality, common in search models, exists because the benefits of search are divided between the searcher and the agent with whom he ultimately trades. Since the searcher (here the advertiser) bears the entire cost of search, he tends to search less than is socially optimal. A second externality occurs because a firm that advertises a low price draws sales and profits away from higher-priced firms. This "undercutting" externality tends to cause too much advertising. The exact cancellation of the two externalities in Butters's model is surprising. Grossman and Shapiro (1984) conjecture that the undercutting externality tends to dominate the search externality in homogeneous goods markets, causing excess advertising. The present paper shows, however, that the search externality generally dominates the undercutting externality, causing every firm to advertise less than is socially optimal. Butters's model is a limiting case of the models constructed in the present paper.

The paper is organized as follows: Section I shows that firms advertise less than is socially optimal in a simple model; Section II generalizes the model in several ways, and underadvertising persists in all cases; Section III shows that the availability of several advertising media may cause excess advertising in some media; Section IV summarizes related research into models of large firms and differentiated goods. All of the analysis is partial equilibrium, meaning that outside goods are implicitly assumed to be priced at marginal cost.

I. The Simple Model

The model in this section is similar to Butters's (1977) model, except that consumers have heterogeneous reservation values. A market opens once. On the supply side of the market, an unlimited quantity of firms produce a homogeneous good at constant unit cost c ($c \geq 0$). On the demand side, small (i.e., infinitesimal) consumers each want to buy one unit of the good. The "reservation value" of a consumer is the maximum price that he is willing to pay for the good; m is the maximum reservation value, and $\tau(v)$ is the fraction of consumers having reservation values less than or equal to v . Assume $m > c$, $\tau(c) = 0$, $\tau(v) < 1$ for $v < m$, and $\tau(v) = 1$ for $v \geq m$. Also assume that τ is continuous.

At the start of the (single) period, each firm must decide what price to advertise and how much advertising to send. A firm that advertises price p is called "firm p ." Firm p 's advertising simply announces that it sells the product at price p . The firm is subsequently required to honor its advertised price.⁵

To motivate the advertising technology, consider first a finite economy such that firms send discrete "ads," the demand side comprises n discrete consumers, and each consumer has $1/n$ chance of observing any given ad. Suppose that advertising is measured in "units" of ads per consumer. For any given block of A units of advertising (i.e., An ads), any given consumer observes none of the ads in that block with probability $(1 - 1/n)^{An}$. As n diverges to infinity, that probability converges to e^{-A} .

Advertising in the infinite economy is measured in "units" analogous to "ads per consumer." For any given block of A units of advertising, assume that any given consumer observes none of the ads in that block with probability e^{-A} . The cost of sending advertising is b per unit ($b > 0$). In equilibrium, firms collectively send a finite

⁵Arguments for honesty include loss of reputation and legal sanctions, but I include no explicit mechanism here.

positive quantity of advertising, but each individual firm sends zero units of advertising.⁶

Firm behavior is completely described by an advertising distribution $\alpha: [c, m] \rightarrow \mathbb{R}_+$; $\alpha(p)$ equals the units of advertising sent by firms charging prices less than or equal to p . After all ads have been sent, each consumer buys one unit of the good from the firm that sent the ad offering the lowest price among all the ads that the consumer observes, if that price does not exceed the consumer's reservation value. A consumer who observes no ads buys nothing. That completes the model. The exogenous parameters are b, c, m , and τ .

An equilibrium is defined to be a continuous advertising distribution $\alpha: [c, m] \rightarrow \mathbb{R}_+$, such that the return to the marginal ad is nonpositive at all prices and zero at every advertised price.⁷ To make that definition precise, let α denote an arbitrary continuous advertising distribution. Suppose that a firm p sends A additional units of advertising (creating a jump in the advertising distribution at price p). The additional units of advertising earn net profits

$$(1) \quad \pi(A; p, \alpha) \\ = (1 - e^{-A})e^{-\alpha(p)}[1 - \tau(p)] \\ \times (p - c) - bA.$$

The first term is the fraction of consumers who observe the new ads. The second term

is the fraction who do not observe lower-priced ads. The third term is the fraction who are willing to pay price p . The fourth term is the gross profit per unit sold. The product of the first four terms is the gross profit generated by the new ads. The last term is the cost of the ads. Given the advertising distribution α , the net profit from the marginal unit of advertising at price p is therefore

$$(2) \quad \mu(p; \alpha) \equiv \frac{\partial \pi(A; p, \alpha)}{\partial A} \Big|_{A=0} \\ = e^{-\alpha(p)}[1 - \tau(p)](p - c) - b.$$

Definition 1: $\alpha: [c, m] \rightarrow \mathbb{R}_+$, a nondecreasing function, is an equilibrium if

- (3a) $\alpha(c) = 0$
- (3b) $\mu(p; \alpha) \leq 0$ for all $p \in [c, m]$
- (3c) $\alpha(p_1) = \alpha(p_2)$
if $\mu(p; \alpha) < 0$ for all $p \in [p_1, p_2]$
- (3d) α is continuous.

The easiest way to show that an equilibrium exists is by construction. Let

$$(4) \quad \beta(c) \equiv 0 \\ \beta(p) \equiv \log \left(\frac{[1 - \tau(p)](p - c)}{b} \right) \\ \text{for } p \in (c, m].$$

Note that $\mu(p; \beta) = 0$ for all $p \in (c, m]$. Let

$$(5) \quad \gamma(p) \equiv \max_{s \in [c, p]} \beta(s) \quad \text{for } p \in [c, m].$$

Obviously, $\gamma: [c, m] \rightarrow \mathbb{R}$ is nondecreasing and satisfies (3a). For all $p \in [c, m]$, $\gamma(p) \geq \beta(p)$ implies $\mu(p; \gamma) \leq 0$, so γ satisfies (3b). Since γ is flat wherever $\gamma(p) \neq \beta(p)$ and is therefore flat wherever $\mu(p; \gamma) \neq 0$, γ satisfies (3c). Also, γ is flat in the neighborhood of c , and β is continuous everywhere else,

⁶Butters (1977) assumes that firms and consumers are finite in number and obtains all of his results in the limit as the numbers diverge to infinity. This approach requires some complicated technical arguments which Butters confines to an appendix. I avoid these complications by assuming from the start that firms and consumers are small. This implies that individual firms do not send positive quantities of advertising. One could assign a positive *density* of advertising to individual firms, but that is an inconvenient construction, because the density of advertising at any given price is indeterminate in equilibrium.

⁷The arguments of Butters (1977) and Kenneth Burdett and Kenneth Judd (1983) show that any equilibrium advertising distribution must be continuous. I avoid an elaborate restatement of those arguments and simply confine attention to continuous distributions.

so γ satisfies (3d). Therefore, γ is an equilibrium; γ happens to be the unique equilibrium, but that is not important for present purposes.

For an example of equilibrium, suppose that reservation values are distributed uniformly on the $[0, 1]$ interval and the unit cost of production is $c = 0$. If information were perfect, every consumer would purchase the good at the competitive price $p = 0$. Instead, suppose that firms must announce their prices through advertising and the unit cost of advertising is $b = 0.1$.⁸ Then, it is straightforward to show [using (5)] that firms advertise prices ranging from 0.11 to 0.50. Forty-eight percent of the consumers purchase the good, at an average price of 0.19.

The equilibrium γ has the curious feature that a mean-preserving spread of consumers' reservation values sometimes *shrinks* the range of advertised prices. For instance, if reservation prices are distributed uniformly on an interval $(M - \varepsilon, M + \varepsilon)$, where $M > c + b$ and ε is small, then (4) implies that β attains a maximum at $p = M - \varepsilon$, and (5) implies that advertising occurs at prices $p \in (c + b, M - \varepsilon)$. The range of advertised prices $(c + b, M - \varepsilon)$ is thus disjoint of the range of reservation values $(M - \varepsilon, M + \varepsilon)$. If ε increases slightly, then the interval of reservation values expands, forcing the interval of advertised prices to contract. It would be interesting to discover whether the inverse relationship between the spread of reservation values and the spread of prices is an

artifact of the example or a more general phenomenon, but that is beyond the scope of this paper.

For any advertising distribution $\alpha: [c, m] \rightarrow \mathbb{R}_+$ (continuous or not, equilibrium or not), the assumption that every consumer accepts the best offer at or below his reservation value implies that total welfare (i.e., surplus) equals

(6) $W(\alpha)$

$$\equiv \int_c^m (v - c)(1 - e^{-\alpha(v)}) d\tau(v) - b\alpha(m).$$

It is not immediately obvious whether firms advertise more or less than is socially optimal. Firms' inability to capture all consumer surplus (the search externality) tends to cause underadvertising, but the possibility of capturing profits from higher-priced firms (the undercutting externality) tends to cause excess advertising. Nevertheless, the search externality *always* dominates the undercutting externality. Increasing advertising at any advertised price increases welfare.

PROPOSITION 1: *Let α be an equilibrium advertising distribution such that $\alpha(m) > 0$. Let q denote the highest advertised price (i.e., q satisfies $\alpha(p) < \alpha(q) = \alpha(m)$ for all $p < q$).⁹ Let $h: [c, m] \rightarrow \mathbb{R}$ be any nondecreasing function such that $h(m) = h(q) > 0$. Then $W(\alpha + \varepsilon h) > W(\alpha)$ for sufficiently small $\varepsilon > 0$.*

PROOF:

First it shall be shown that $\mu(q; \alpha) = 0$ and $q < m$. If $\mu(q; \alpha) < 0$, then (3d) implies $\mu(p; \alpha) < 0$ for all $p \in [q - \delta, q]$ for some $\delta > 0$. Condition (3c) then implies $\alpha(q - \delta) = \alpha(q)$, a contradiction. Therefore, $\mu(q; \alpha) = 0$, and (2) implies $q < m$. Equation (6)

⁸This footnote shows that b can be interpreted as the marginal cost of reaching one additional consumer through advertising. Suppose that b is the cost of one unit of advertising in the n -consumer economy used to motivate the advertising technology. Since each unit of advertising comprises n ads, the marginal cost of reaching one more consumer is b/n . If each consumer's potential purchase is normalized to $1/n$, then the marginal cost of reaching $1/n$ more potential purchases is b/n . In the limit, the marginal cost of reaching potential purchases is b per potential purchase. If consumers are now renormalized so that each consumer makes one potential purchase, then b is the marginal cost of reaching one additional consumer through advertising.

⁹It is impossible to determine whether firm q actually exists (cf. footnote 6), but for expositional purposes it is convenient to assume that firm q does exist. The informal arguments based on the existence of firm q can all be reposed in terms of firms advertising prices in the neighborhood of q , which certainly do exist.

implies

$$\begin{aligned}
 & W(\alpha + \varepsilon h) \\
 &= \int_c^m (v - c)[1 - e^{-\alpha(v) - \varepsilon h(v)}] d\tau(v) \\
 &\quad - b\alpha(m) - b\varepsilon h(q) \\
 &\frac{\partial W(\alpha + \varepsilon h)}{\partial \varepsilon} \bigg|_{\varepsilon=0} \\
 &= \int_c^m (v - c)h(v) \cdot e^{-\alpha(v)} d\tau(v) - bh(q) \\
 &\geq \int_q^m (v - c)h(q) \cdot e^{-\alpha(q)} d\tau(v) - bh(q) \\
 &> h(q) \cdot \left[\int_q^m (q - c)e^{-\alpha(q)} d\tau(v) - b \right] \\
 &= h(q) \cdot \{(q - c)e^{-\alpha(q)}[1 - \tau(q)] - b\} \\
 &= h(q) \cdot \mu(q; \alpha) = 0.
 \end{aligned}$$

The intuition is as follows. Ads at prices in the neighborhood of m have a vanishingly small probability of being observed by a consumer who is willing to pay that price, so firms do not advertise prices in the neighborhood of m , implying $q < m$. Firm q captures no sales from other firms, and $q < m$ implies that it does not capture all the surplus from new sales it creates. Therefore, firm q has a socially inadequate incentive to advertise and could increase welfare by increasing its advertising. Therefore, any firm $p < q$ could increase welfare by increasing its advertising, because additional ads at prices $p < q$ increase total social surplus at least as much as additional ads at price q .

This line of reasoning avoids the problem of comparing the externalities directly. Only the search externality applies to the highest-priced firm, and the balance of the externalities for lower-priced firms is inferred indirectly from the argument that additional advertising at low prices must be

at least as good as additional advertising at high prices.¹⁰

Proposition 1 shows that every firm p can increase welfare by buying more advertising, but it does not follow that more advertising at every price p eventually leads to a globally optimal advertising distribution. A global optimum generally requires shifting advertising from high prices to low prices. Proposition 2 describes a global optimum in which c is the only advertised price. Firms lose money, but surplus is maximized because, for any given quantity of advertising, it is socially optimal to set the price of the good equal to its marginal cost of production. Proposition 2 shows that a globally optimal advertising distribution is $\eta: [c, m] \rightarrow \mathbb{R}$, where

$$(7) \quad \eta(p) \equiv \max(0, \log[(\bar{v} - c)/b])$$

for all $p \in [c, m]$

$$\bar{v} \equiv \int_c^m v d\tau(v)$$

(\bar{v} is consumers' mean reservation value).

PROPOSITION 2: $W(\eta) \geq W(\alpha)$ for all nondecreasing $\alpha: [c, m] \rightarrow \mathbb{R}_+$.

PROOF:

For any $A \geq 0$, let $\eta_A: [c, m] \rightarrow \mathbb{R}_+$ denote the ad distribution such that $\eta_A(p) = A$ for all $p \in [c, m]$. Equation (6) implies

$$W(\eta_A) = (1 - e^{-A})(\bar{v} - c) - bA.$$

¹⁰ It is not difficult to see at this point why every firm advertises exactly optimally in Butters's (1977) model. In Butters's model, all consumers have a common reservation value m , which must be the highest advertised price, since it is clearly suboptimal for the highest-priced firm to charge a lower price. Therefore, neither externality applies to the highest-priced firm, which consequently advertises exactly optimally. Since the marginal ad at any price below m creates the same number of aggregate sales and makes the same contribution to welfare, the marginal social return to ads at any price below m is also zero.

This function is strictly concave in A and is maximized at $A = \log[(\bar{v} - c)/b]$. For any nondecreasing $\alpha: [c, m] \rightarrow \mathbb{R}_+$, (6) implies

$$\begin{aligned} W(\eta_{\alpha(m)}) - W(\alpha) \\ = \int_c^m (v - c)[e^{-\alpha(v)} - e^{-\alpha(m)}] d\tau(v) \\ \geq 0. \end{aligned}$$

Therefore, $W(n) \geq W(\eta_{\alpha(m)}) \geq W(\alpha)$.

The welfare loss from suboptimal advertising is difficult to calculate. If reservation values are distributed uniformly on $[c, m]$, then I have shown in Stegeman (1986) that $W(\eta) - W(\alpha) < b/2$, where α denotes any equilibrium advertising distribution. $W(\eta) - W(\alpha)$ can be interpreted as the average welfare loss per consumer, and b can be interpreted as the average cost of reaching one consumer through advertising (cf. footnote 8).

II. Generalizations

This section generalizes the model in several ways. Every generalization preserves the competitive assumption that small agents are trading a homogeneous good, and every generalization but one preserves the assumption that all price information flows through a single advertising medium. Within these bounds, the underadvertising result is quite robust.

The first generalization is trivial. If firms have generally heterogeneous and nondecreasing marginal cost curves, then free entry implies that only firms producing at the lowest possible cost advertise in equilibrium; c denotes that cost.

A. Directed Advertising

In Section I, every consumer was equally likely to observe ads. In the present section, I divide consumers into a continuum of types, such that consumer's type determines his propensity to receive advertising and may be correlated with his reservation value.

It is still true that every firm underadvertises. The next two paragraphs sketch the argument. An interesting by-product of this analysis is the construction of a large family of advertising technologies, which may be useful in other applications.

Assume that consumers are divided into M types, where some types are more likely to observe advertising than others. Let j_m denote the fraction of consumers of type m ($\sum_{m=1}^M j_m = 1$). To set up the limit argument used in Section I, suppose for the moment that there are only n consumers and that each consumer of type m has k_m/n chance of observing any given ad. If a firm (or firms) sends A "ads per consumer" (i.e., An ads), then, as n diverges to infinity, the probability that a given consumer of type m observes none of those ads converges to $e^{-k_m A}$. The total fraction of consumers who observe at least one ad converges to

$$\begin{aligned} (8) \quad \psi(A) &\equiv 1 - j_1 e^{-k_1 A} - j_2 e^{-k_2 A} \dots \\ &\quad - j_M e^{-k_M A}. \end{aligned}$$

If M is large, then ψ can take a wide variety of shapes; each ψ can be interpreted as a distinct advertising technology. Butters's (1977) technology (also used in Section I) is the special case $\psi(A) = 1 - e^{-A}$. In every case, ψ is differentiable, strictly increasing, and strictly concave.

If a consumer's type is correlated with his reservation value, then $\psi(A)$ does not contain enough information to determine the equilibrium conditions. Therefore, $\psi(A)$ must be generalized to $\psi(v, A)$, where $\psi(v, A)$ denotes the fraction of consumers having reservation value v who observe at least one of the ads in any given block of A units of advertising. Assume that $\psi(v, A)$ is continuous in (v, A) , strictly increasing, strictly concave, and continuously differentiable in A , and that it satisfies $\psi(v, 0) = 0$ for all v . The general advertising technology $\psi(\cdot, \cdot)$ can then be transplanted into the model of Section I by substituting (1'), (2'),

and (6') for (1), (2), and (6).

$$(1') \quad \pi(A; p, \alpha) \equiv \int_p^m [\psi(v, \alpha(p) + A) - \psi(v, \alpha(p))] d\tau(v) \\ \times (p - c) - bA$$

$$(2') \quad \mu(p; \alpha) \\ \equiv \int_p^m \frac{\partial \psi(v, \alpha(p) + A)}{\partial A} \Big|_{A=0} d\tau(v) \\ \times (p - c) - b$$

$$(6') \quad W(\alpha) \equiv \int_c^m (v - c) \psi(v, \alpha(v)) d\tau(v) \\ - b\alpha(m).$$

Given these substitutions, the definition of equilibrium and the statement of Proposition 1 are unchanged. The proof of Proposition 1 is similar to the original proof.¹¹ Every firm advertises less than is socially optimal.

Directed advertising is interesting partly because it generates a large family of advertising technologies corresponding to various specifications of $\psi(v, A)$. For convenience, call this the exponential family. Butters's (1977) technology is just one member of the exponential family. This section has shown that Proposition 1 extends to every member of the exponential family.

Advertising technologies in the exponential family share two special properties. First, they exhibit no economies to scale. This is necessary in competitive models to ensure that firms are small. Second, the formulation of ψ [equation (8)] presupposes that ads are homogeneous. Section III investigates the consequences of heterogeneous ads.

B. Nonrectangular Demand Curves

Every consumer had a rectangular demand curve in section I, but the underadvertising result does not depend on that assumption. The argument follows. Suppose that consumers have generally heterogeneous demands and that total demand is strictly decreasing as a function of price. In equilibrium, some consumers must be willing to pay a price higher than q (the highest advertised price) for some units of the good, because otherwise firm q would lose money (cf. footnote 9). The marginal ad from firm q creates sales to consumers in that group who receive no other ads, but firm q does not capture all of the consumer surplus from such sales. Thus, firm q underadvertises. Therefore every firm underadvertises, since the marginal ad from firm $p < q$ increases total welfare at least as much as the marginal ad from firm q .

C. Communication Among Consumers

The model of Section I assumed that each consumer receives information only from the advertisements he observes. Instead, suppose that consumers share information. If each ad is passed on to several randomly chosen consumers, and each ad is passed on to the same expected number of consumers, then the revised model is equivalent to the original model with a lower cost of advertising b , and Proposition 1 is unchanged.

Alternatively, if lower-priced ads are passed around more than higher-priced ads, then the argument that welfare could be increased by sending more ads at price q is unchanged, and the argument that sending more ads at $p < q$ increases welfare at least as much as sending more ads at price q is reinforced. It is still true that every firm underadvertises.

To summarize, Section II has shown that Proposition 1 is quite robust. The conclusion that firms underadvertise is sensitive neither to the advertising technology nor to the market participants, which may be generally heterogeneous. Underadvertising also

¹¹The assumed properties of ψ ensure the continuity of μ in p , which implies $q < m$. The rest of the proof is a direct extension of the proof of Proposition 1, if one remembers that $\alpha(p) = \alpha(q)$ for all $p \in [q, m]$.

extends to models incorporating certain kinds of information exchange among consumers.

III. Several Advertising Media

In Section I, ads were homogeneous, in the sense that any given consumer was equally likely to observe any ad. This section describes a market in which firms can advertise in several different media, which tend to reach different classes of consumers. One medium might be particularly good for reaching consumers with high reservation values, and another might be good for reaching consumers with low reservation values. It will be shown that underadvertising must occur in some media but that excess advertising may occur in others.

An equilibrium in a model of M media is (extending the definition of Section I) a set of continuous advertising distributions, $\alpha_i: [c, m] \rightarrow \mathbb{R}_+$, $i = 1, 2, \dots, M$, such that the return to the marginal ad in medium i is nonpositive everywhere and zero at any price at which firms advertise in medium i , for all i . A more formal definition is not essential for most of what follows. It is useful, however, to distinguish two kinds of media in equilibrium. For a given equilibrium, let q_i denote the highest price charged in medium i [i.e., q_i satisfies $\alpha_i(p) < \alpha_i(q_i) = \alpha_i(m)$ for all $p < q_i$], and let $q \equiv \max_i q_i$ be the highest price advertised (cf. footnote 9). The medium or media used by the highest-priced firms (i.e., the media i such that $q_i = q$) are *full-price* media. Other media are *discount* media.¹²

A. Full-Price Media

The underadvertising result extends to full-price media. In other words, every firm underadvertises in every full-price medium.

The reasoning is similar (once again) to the reasoning behind the original Proposition 1. Firm q advertises only in full-price media and must be advertising less than is socially optimal in those media, because the marginal ad from firm q takes no sales from other firms and captures only part of the surplus from sales it creates. Any firm $p < q$ must also be advertising less than is socially optimal in every full-price medium, because ads at prices $p < q$ create at least as many new sales as ads at price q (in the same medium). Therefore, every firm underadvertises in every full-price medium.

The extension of Proposition 1 to all full-price media has interesting implications. For example, suppose that all advertising occurs in newspapers. If different consumers read different newspapers and advertising rates vary by newspaper, then distinct newspapers must be considered distinct media, and Proposition 1 does not apply. If, however, the highest-priced firms advertise in every newspaper, then all of the media are full-price media, implying that every firm underadvertises in every newspaper.

B. Excess Advertising in Discount Media

This section shows that Proposition 1 does not extend to discount media, by constructing an example of excess advertising in a discount medium. The example has the following general features. Firms can advertise in two media, D and F. Medium D is better than medium F for reaching low-reservation-value consumers, and medium F is better for reaching high-reservation-value consumers. In equilibrium, medium D becomes a discount medium, and medium F becomes a full-price medium. Medium D also reaches some high-reservation-value consumers and, in particular, reaches those who are also most likely to observe ads in medium F. The firms advertising the highest prices in medium D thus capture some sales from higher-priced firms advertising in medium F, and this negative externality is so strong that the former firms advertise too much in medium D.

To make the point more concrete, suppose that reservation values are correlated

¹² Whether a medium is a full-price or discount medium is determined in equilibrium. If there were multiple equilibria, then it is conceivable that a discount medium in one equilibrium could be a full-price medium in another equilibrium.

with wealth and that the *Times* is relatively better for reaching wealthy consumers than the *Post*. The essential characteristic of the example is that there exists a positive correlation between *Times* readers and *Post* readers among wealthy consumers, meaning that the wealthy population tends to divide into readers versus nonreaders rather than into *Times* readers versus *Post* readers. Then the firms advertising the highest prices in the *Post* capture so many sales from even-higher-priced firms advertising in the *Times* that the marginal social product of their advertising is negative.

The remainder of this section describes the particulars of the example. Assume that all firms produce the good at constant unit cost c ($c \geq 0$). Consumers are divided equally into three types. Type s has reservation value s , and types m_1 and m_2 have reservation value m ($m > s > c$). Firms have access to two advertising media (D and F). A consumer of type s has $1 - e^{-A}$ chance of observing at least one ad in any given block of A ads in medium D and does not observe ads in medium F. A consumer of type m_1 does not observe ads in medium D and has $1 - e^{-kA}$ chance of observing at least one ad in any given block of A ads in medium F, where k ($0 < k < 1$) is exogenous. A consumer of type m_2 has $1 - e^{-A}$ chance of observing at least one ad in any given block of A ads, regardless of the medium. Type m_1 consumers are (relative) nonreaders; type m_2 consumers are readers. The cost of advertising in either medium is b per unit ($b > 0$). The timing of the model is the same as in Section I. Each firm chooses a price and decides how much advertising to send in each medium. After all advertising has been sent, each consumer purchases one unit of the good from the firm charging the lowest price among the ads that he observes, if that price does not exceed his reservation value. The exogenous parameters of the example are b , c , s , m , and k . Assume that $s > c + b/2k$.

The next paragraphs provide a precise definition of equilibrium for the two-media model, describe a particular equilibrium (ϕ, δ) , define welfare for the two-media model, and show that excess advertising oc-

curs in medium D in equilibrium (ϕ, δ) . The definition of equilibrium is similar to Definition 1. Let α_D and α_F denote continuous advertising distributions in media D and F, respectively. A additional units of advertising at price p in medium D yield net additional profits

$$\begin{aligned} (9) \quad \pi_D(A; p, \alpha_D, \alpha_F) \\ &= (1 - e^{-A})e^{-\alpha_D(p)}(p - c) \\ &\quad + (1 - e^{-A})e^{-\alpha_D(p) - \alpha_F(p)} \\ &\quad \times (p - c) - bA. \end{aligned}$$

The first term is profit from sales to consumers of type s . The second term is profit from sales to consumers of type m_2 . A additional units of advertising at price p in medium F yield net profits

$$\begin{aligned} (10) \quad \pi_F(A; p, \alpha_D, \alpha_F) \\ &= (1 - e^{-kA})e^{-k\alpha_F(p)}(p - c) \\ &\quad + (1 - e^{-A})e^{-\alpha_D(p) - \alpha_F(p)} \\ &\quad \times (p - c) - bA. \end{aligned}$$

The first term is profit from sales to consumers of type m_1 . The second term is profit from sales to consumers of type m_2 . The expected profits from the *marginal* ad at price p [cf. (2)] in media D and F, respectively, are

$$\begin{aligned} (11a) \quad \mu_D(p; \alpha_D, \alpha_F) \\ &= (e^{-\alpha_D(p)} + e^{-\alpha_D(p) - \alpha_F(p)}) \\ &\quad \times (p - c) - b \\ (11b) \quad \mu_F(p; \alpha_D, \alpha_F) \\ &= (ke^{-k\alpha_F(p)} + e^{-\alpha_D(p) - \alpha_F(p)}) \\ &\quad \times (p - c) - b. \end{aligned}$$

Firms charging prices $p > s$ obviously do not use medium D in equilibrium, because medium F is just as good for reaching type-

m_2 consumers and better for reaching type- m_1 consumers. The following definition of equilibrium incorporates this observation.

Definition 2: An equilibrium (in the two-media example) is a pair of nondecreasing functions, $\alpha_D: [c, s] \rightarrow \mathbb{R}_+$ and $\alpha_F: [c, m] \rightarrow \mathbb{R}_+$, that satisfy (12a)–(12d) for $i = D, F$.

$$(12a) \quad \alpha_i(c) = 0$$

$$(12b) \quad \mu_D(p; \alpha_D, \alpha_F) \leq 0 \quad \text{for all } p \in [c, s]$$

$$\mu_F(p; \alpha_D, \alpha_F) \leq 0 \quad \text{for all } p \in [c, m]$$

$$(12c) \quad \alpha_i(p_1) = \alpha_i(p_2)$$

$$\text{if } \mu_i(p; \alpha_D, \alpha_F) < 0 \quad \text{for all } p \in [p_1, p_2]$$

$$(12d) \quad \alpha_i \text{ is continuous.}$$

Lemma 1 describes a particular equilibrium. In this equilibrium, firms charging prices in the interval $[c + b/2, s]$ advertise in medium D and firms charging prices in the overlapping interval $[c + b/2k, m]$ advertise in medium F. The advertising distribution for medium F is ϕ , where

$$(13) \quad \phi(p) \equiv 0 \quad \text{for } p \in [c, c + b/2k]$$

$$\phi(p) \equiv f^{-1}(0) \quad \text{for } p \in (c + b/2k, s]$$

$$\phi(p) \equiv g^{-1}(0) \quad \text{for } p \in (s, m]$$

where

$$f(x) \equiv ke^{-kx}(1 + e^{-x})(p - c) - b$$

and

$$g(x) \equiv ke^{-kx}(1 + e^{-k\phi(s) - (1-k)x}) \\ \times (p - c) - b.$$

Note that $f^{-1}(0)$ and $g^{-1}(0)$ are well-defined because f and g are monotonically decreasing and map \mathbb{R} onto $(-b, +\infty)$. The distribution of advertising in medium D is

δ , where

$$(14) \quad \delta(p) \equiv 0 \quad \text{for } p \in [c, c + b/2]$$

$$\delta(p) \equiv \log[2(p - c)/b] \\ \text{for } p \in (c + b/2, c + b/2k]$$

$$\delta(p) \equiv k\phi(p) - \log(k) \\ \text{for } p \in (c + b/2k, s].$$

LEMMA 1: (δ, ϕ) is an equilibrium.

PROOF:

It is straightforward to show that δ and ϕ are continuous and nondecreasing and that they satisfy $\delta(c) = \phi(c) = 0$. Direct substitution shows: $\mu_D(p; \delta, \phi) \leq 0$ for all $p \in [c, c + b/2]$; $\mu_D(p; \delta, \phi) = 0$ for all $p \in [c + b/2, s]$; $\mu_F(p; \delta, \phi) \leq 0$ for all $p \in [c, c + b/2k]$; $\mu_F(p; \delta, \phi) = 0$ for all $p \in [c + b/2k, m]$. Therefore, (δ, ϕ) also satisfies (12b) and (12c).

The next step is to define welfare in the two-media model. For any given pair of advertising distributions α_D and α_F , the assumption that every consumer accepts the best offer at or below his reservation value implies that total welfare (i.e., surplus) equals

$$(15) \quad W(\alpha_D, \alpha_F) \\ \equiv (1 - e^{-\alpha_D(s)})(s - c) \\ + (1 - e^{-k\alpha_F(m)})(m - c) \\ + (1 - e^{-\alpha_D(s) - \alpha_F(m)})(m - c) \\ - b[\alpha_D(s) + \alpha_F(m)].$$

The first three terms are the surplus from sales to consumers of types s , m_1 , and m_2 , respectively, and the fourth term is the total cost of advertising.

Proposition 3 shows that firms advertise too much in medium D in equilibrium (ϕ, δ) .

PROPOSITION 3: Let $h: [c, s] \rightarrow \mathbb{R}_+$ be any nondecreasing function such that $h(c) = 0$, $h(s) > 0$, and $\delta - h$ is nondecreasing. Then, $W(\delta - \varepsilon h, \phi) > W(\delta, \phi)$ for sufficiently small $\varepsilon > 0$.

PROOF:

The proof of Lemma 1 shows that $\mu_D(s; \delta, \phi) = \mu_F(s; \delta, \phi) = \mu_F(m; \delta, \phi) = 0$. Equations (11) imply

$$\begin{aligned} (16a) \quad \mu_D(s; \delta, \phi) &= (e^{-\delta(s)} + e^{-\delta(s)-\phi(s)})(s-c) - b \\ &= 0 \end{aligned}$$

$$\begin{aligned} (16b) \quad \mu_F(s; \delta, \phi) &= (ke^{-k\phi(s)} + e^{-\delta(s)-\phi(s)})(s-c) - b \\ &= 0 \end{aligned}$$

$$\begin{aligned} (16c) \quad \mu_F(m; \delta, \phi) &= (ke^{-k\phi(m)} + e^{-\delta(s)-\phi(m)})(m-c) - b \\ &= 0. \end{aligned}$$

Let $\hat{h} \equiv h/h(s)$. Then (15) implies

$$\begin{aligned} (17) \quad \partial W(\delta - \varepsilon \hat{h}, \phi) / \partial \varepsilon|_{\varepsilon=0} &= b - (s-c)e^{-\delta(s)} \\ &\quad - (m-c)e^{-\delta(s)-\phi(m)}. \end{aligned}$$

Substituting (16a) into (17) yields

$$\begin{aligned} (18) \quad \partial W(\delta - \varepsilon \hat{h}, \phi) / \partial \varepsilon|_{\varepsilon=0} &= (s-c)e^{-\delta(s)-\phi(s)} \\ &\quad - (m-c)e^{-\delta(s)-\phi(m)}. \end{aligned}$$

Equations (16b) and (16c) imply

$$(19) \quad s-c = (m-c) \frac{ke^{-k\phi(m)} + e^{-\delta(s)-\phi(m)}}{ke^{-k\phi(s)} + e^{-\delta(s)-\phi(s)}}.$$

Substituting (19) into (18) yields, after rou-

tine factorization,

$$\begin{aligned} \partial W(\delta - \varepsilon \hat{h}, \phi) / \partial \varepsilon|_{\varepsilon=0} &= \frac{k(m-c)e^{-\delta(s)-\phi(m)-k\phi(s)}}{ke^{-k\phi(s)} + e^{-\phi(s)-\delta(s)}} \\ &\quad \times [e^{(1-k)(\phi(m)-\phi(s))} - 1] > 0. \end{aligned}$$

That establishes the result.

The coincidence of excess advertising in the discount medium and underadvertising in the full-price medium is slightly paradoxical, because it has been emphasized repeatedly that (at the margin) low-price ads increase welfare more than high-price ads. That is true within any single advertising medium, but not across media.¹³ In the example, the marginal highest-priced ad increases welfare more than the marginal lowest-priced ad because the favorable effect of advertising in the full-price medium overwhelms the unfavorable effect of charging the full price.

It is important to emphasize that there is no *presumption* that firms advertise excessively in discount media. The above example only demonstrates the *possibility*. It is easy to construct examples in which excess advertising does not occur (e.g., the above model, with $k \geq 1$). Indeed, positive correlation of readerships is implausible for newspapers.

IV. Related Research

All of the models in this paper satisfy the fundamental competitive assumption that many small firms produce a homogeneous good. One way to relax that assumption is

¹³In this particular example, firms that charge prices $p \in [s, m]$ advertise exactly optimally, but this is an artifact of the discontinuous distribution of reservation values (cf. footnote 10). If the parameters of the example are perturbed slightly to smooth the distribution of reservation values, then the argument of Section III-A implies that all firms advertise too little in medium F, and it is still true that all firms that advertise in medium D advertise too much in medium D.

to introduce large firms. Stahl (1989) constructs a model similar to the present model, except that firms enjoy economies of scale in advertising. In a one-shot symmetric mixed-strategy equilibrium, all firms under-advertise. Another way to relax the competitive assumption is to introduce differentiated goods. Elsewhere (Stegeman, 1986) I construct a model that is similar to the simple model of Section I, except that the product space is a circle and each firm chooses its product as well as its price and quantity of advertising. Small consumers are distributed uniformly around the circle. After firms send advertising, each consumer responds to the best offer that he receives, taking into account price and linear transportation costs. It is shown that firms under-advertise in any symmetric equilibrium, except for a borderline case in which firms advertise exactly optimally. The robustness of these results is unclear. Grossman and Shapiro (1984) construct a model in which firms exploiting economies of scale in advertising distribute themselves around the circle, and they find excess advertising in some cases.¹⁴

A different way to generalize the model is to admit the possibility of consumer search. Butters (1977) describes a bilateral-search model in which consumers search after observing ads. Butters shows that the addition of consumer search causes firms to advertise too much. Elsewhere (Stegeman, 1986) I add a similar consumer search process to the model of Section I. In that case, consumer search may or may not cause excess advertising, but every firm does advertise too much if consumer search costs are sufficiently small. A partial explanation is that the firm advertising the highest price makes a substantial fraction of its sales to con-

sumers who would otherwise have purchased after search. All of those sales are captured from lower-priced firms, a negative externality similar to the undercutting externality.

V. Conclusion

The contention that firms advertise too much has provoked much inconclusive debate. The complexity of the issues precludes any simple resolution, but it is possible to obtain a better understanding of the circumstances that tend to lead to excess advertising. This paper contributes to such an understanding. The general setting is a competitive market (i.e., small agents, free entry, and a homogeneous good) in which the sole purpose of advertising is to convey price information. In this setting, it has been shown that firms buy less price advertising than is socially optimal. The result is quite robust with respect to the agents in the market and the advertising technology. If firms advertise in several media, however, then an example has shown that excess advertising can occur in media used exclusively to advertise discount prices.

The conclusion that competitive firms provide too little informative advertising complements the conclusion of Stigler and Becker (1977) and Nichols (1985) that competitive firms provide the optimal quantity of persuasive advertising. These results jointly establish a foundation for the welfare analysis of advertising in noncompetitive markets.

REFERENCES

- Beales, Howard, Craswell, Richard and Salop, Steven, "The Efficient Regulation of Consumer Information," *Journal of Law and Economics*, December 1981, 24, 491-539.
- Burdett, Kenneth and Judd, Kenneth, "Equilibrium Price Dispersion," *Econometrica*, July 1983, 51, 955-70.
- Butters, Gerard, "Equilibrium Distributions of Sales and Advertising Prices," *Review of Economic Studies*, October 1977, 44, 465-91.
- Grossman, Gene, M. and Shapiro, Carl, "Infor-

¹⁴Grossman and Shapiro (1984) make the general claim that advertising in their model always exceeds the social optimum if equilibrium sales S are sufficiently high, but their demonstration is unconvincing. They essentially show the existence of S^* such that $S > S^*$ implies excessive advertising but do not point out that S^* depends upon the parameters of the model. It is not obvious from their general analysis that there exists any set of parameters x such that the $S(x) > S^*(x)$.

- mative Advertising with Differentiated Products," *Review of Economic Studies* January 1984, 51, 63-84.
- Kaldor, Nicholas, "The Economic Aspects of Advertising," *Review of Economic Studies*, January 1950, 18, 1-27.
- Nichols, Len. M., "Advertising and Economic Welfare," *American Economic Review*, March 1985, 75, 213-8.
- Salop, Steven, "Monopolistic Competition with Outside Goods," *Bell Journal of Economics*, February 1979, 10, 141-56.
- Schmalensee, Richard, *The Economics of Advertising*, Amsterdam: North Holland, 1972.
- Stahl, Dale O., "Oligopolistic Pricing and Advertising," CER Working Paper 89-08, University of Texas at Austin, 1989.
- Stegeman, Mark, "Essays in Search and Speculation," Ph.D. Dissertation, Cambridge, MA: Massachusetts Institute of Technology, 1986.
- _____, "Limit Equilibria of a Bilateral Search Model," mimeo, University of North Carolina, 1990.
- Steiner, Peter O., "Economics of Broadcasting and Advertising—Discussion," *American Economic Review*, May 1966, 56 (*Papers and Proceedings*), 472-5.
- Stigler, George and Becker, Gary, "De Gustibus Non Est Disputandum," *American Economic Review*, March 1977, 67, 76-100.
- Telser, Lester G., "Advertising and Competition," *Journal of Political Economy*, December 1964, 72, 537-62.
- Worcester, Dean A., *Welfare Gains from Advertising*, Washington, DC: American Enterprise Institute, 1976.

High and Declining Prices Signal Product Quality

By KYLE BAGWELL AND MICHAEL H. RIORDAN *

High and declining prices signal a high-quality product. High prices are the efficient means of signaling, because the consequent loss of sales volume is most damaging for lower-cost, lower-quality products. As time passes, and the number of informed consumers increases, the signaling distortion lessens, resulting in a declining price profile. The prediction of high and declining prices is robust across a variety of dynamic models and is consistent with recent empirical findings. (JEL 022, 026)

The marketing literature has produced various evidence on price-quality relationships. Numerous experimental studies show that consumers infer a higher quality from a higher price (Kent B. Monroe, 1973). This inference is consistent with the findings of several case studies. Such diverse products as fountain-pen ink and car wax (André Gabor and Clive Granger, 1965) and vodka, skis, and television sets (Robert D. Buzzell et al., 1972) have been successfully introduced at high prices to connote high quality. A variety of empirical data is also available. Analyses of *Consumer Reports* data yield positive price-quality rank-order correlations for many products, and particularly for consumer durables (Eitan Gerstner, 1985; Gerard J. Tellis and Birger Wernerfeld, 1987). Moreover, a recent longitudinal analysis of *Consumer Reports* data for consumer durables indicates declining trends in (a) real prices, (b) price differentials between competing brands, and (c) the rank-order correlation between price and

quality (David J. Curry and Peter C. Riesz, 1988).¹

These "stylized facts" are consistent with two important features of markets. First, firms signal high-quality new products with prices that are above full-information profit-maximizing prices. Second, over time, as information about the product diffuses, this price distortion lessens or vanishes entirely.

We demonstrate the logic and robustness of this argument in several equilibrium models of behavior by consumers and firms. The models have different assumptions about consumer information. However, all of the models possess intuitively plausible equilibria in which higher-quality products are introduced at higher prices that decline over time.

Our essential argument is outlined as follows. Consider a market in which a firm introduces a new product possessing some innovative feature of uncertain quality. Some consumers can ascertain the quality, while others cannot, but all understand that a higher-quality product is more costly to produce. The most efficient way for the firm to signal high quality is to charge a price too high to be profitable if the product were in fact of lower quality. This high-price strat-

*Department of Economics, Northwestern University, Evanston, IL 60208, and Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215, respectively. We thank Scott Davis, Ray Deneckere, Garey Ramey, Birger Wernerfeld, two anonymous referees, and seminar participants at the Universities of California at Berkeley and Los Angeles, the University of Chicago, Northwestern University, and the 1987 Winter AEA Meeting for helpful comments. This research has been supported by the National Science Foundation through grant numbers IST-8507300 and IRI-8706150.

¹Curry and Riesz collected all test-study evaluations with five or more brands reported in *Consumer Reports* over a 20-year period. They retained only those durable products whose characteristics remained stable during at least three separate publication time periods.

egy is potentially successful for two reasons. First, the consequent loss of sales volume is less damaging to a higher-cost product. Second, a lower-quality product would lose more sales from informed consumers by charging a high price. Understanding this, uninformed consumers rationally infer higher quality from the higher price.

However, as consumers gain experience with the product and information about its quality diffuses, the portion of uninformed consumers in the market declines. Consequently, it becomes even more costly for the firm to signal a higher quality falsely to the uninformed. The firm can efficiently signal a higher quality with a smaller price distortion. Thus, a high and declining price path identifies a high-quality product.

A positive correlation between price and quality follows, because higher-quality products are more costly to produce, so that signaling distorts upward the price of newly introduced high-quality products. As information diffuses, signaling distortions diminish, and the prices of newer products converge downward to those of older products of corresponding quality. An associated weakening of the correlation between price and quality can then be explained by measurement errors in the data. Thus, the theory appears to be consistent with the stylized facts.

Our conclusion that high-quality products have a downward sloping price profile differs from that of previous theoretical contributions to the economics literature. For example, Carl Shapiro (1983) shows that a monopolist charges a high and declining price if consumers optimistically overestimate product quality and charges a low introductory price if consumers pessimistically underestimate product quality. However, these conclusions depend on an assumption that consumers have adaptive expectations about product quality with no possibility of price signaling.

Paul Milgrom and John Roberts (1986) focus on the introductory phase of a non-durable product's life and argue that prices will rise over time as repeat buyers learn about their own preferences. Their analysis is similar to ours in that they also recognize

the potential for high prices to signal high (expected) quality due to cost effects. However, we abstract from short-run experimental-buying effects and focus on the long-run trends associated with the signaling of product quality.²

In a dynamic model of consumer learning, Kenneth L. Judd and Riordan (1987) show that high-quality prices tend to rise after the introductory period. This is because signaling does not occur until after consumers have gained experience with the product, which follows from an assumption of cost parity for different-quality products. Moreover, they conjecture that, for multi-period extensions, price would eventually decline as consumers gain further experience.

John Conlisk et al. (1984) and Nancy L. Stokey (1981) have argued for a declining price path for a durable good of known quality. This path represents the firm's attempt to "skim" the market of higher-valuation buyers. Edward P. Lazear (1986) also predicts a declining price path, under the assumptions that the firm is unsure of the size of its demand and that consumers know quality. While these theories are complementary to ours, they do not provide direct insight into the relationships between price and quality described above.

Our distinction between informed and uninformed consumers is reminiscent of a related literature on product selection in which some consumers observe quality while others do not. In this context, Yuk-Shee Chan and Hayne Leland (1982), Russell Cooper and Thomas W. Ross (1984, 1985), Scott Davis (1989), Joseph Farrell (1980), Riordan (1986), and Asher Wolinsky (1983) argue that the presence of informed consumers enables high prices to signal high-

²This focus seems justified given the nature of the long-run data we seek to understand (see footnote 1). Further, introductory offers to inspire repeat purchases are less relevant for consumer durables than for non-durables, because of the infrequency of consumer purchases of durable goods. For other models of introductory offers and repeat purchases that focus on different issues, see Bagwell (1987), Jacques Crémer (1984), and Joseph Farrell (1986).

quality choices. While our work is related, we take quality to be exogenous and also analyze the role of production costs in establishing high prices as signals. Indeed, we find that high prices can signal high quality even if all consumers are uninformed. Informed consumers are not necessary for the signaling of a given quality, but they do determine the size of the signaling distortion.

Finally, we note independent work by Garey Ramey (1986) and Doron Fertig (1988). They analyze a static model with a continuum of types and no informed consumers and demonstrate a unique separating equilibrium in which high prices signal quality. Our model has only two cost-quality types, introduces informed consumers, and uses the "intuitive criterion" (In-Koo Cho and David M. Kreps, 1987) to select among equilibria. Ramey and Fertig do not develop predictions about the time path of prices.³

The paper is organized in three sections. Our basic results are developed in a static context in Section I. Various multiperiod extensions are analyzed in Section II. Sections I and II may have methodological interest. The intuitive criterion is actively employed in each section, and the criterion is applied in Section II to dynamic signaling games with the possibility of multiple dimensions of private information. Our conclusions are summarized in Section III.

I. Basic Model

Consider a one-period consumer market in which a firm has introduced a new product with a novel feature of uncertain quality. For simplicity, assume that quality is either high or low: $q \in \{H, L\}$.

The production technology is common knowledge. The average cost of a high-quality product is constant and equal to $c > 0$,

while low-quality production cost is normalized to zero without loss of generality.

There are many potential consumers of the new product, approximated by a continuum of mass M (Judd, 1985), each with a potential demand for one unit. Consumers have a common reservation price, $P^L > 0$, for a low-quality product. On the other hand, consumers have heterogeneous reservation prices for a high-quality product, uniformly distributed between P^L and $(1 + P^L)$. The uniform distribution is convenient because it generates a linear demand for a high-quality product.

Some consumers are informed about product quality, while remaining consumers believe that quality is high with probability r . This prior belief is common knowledge. Let X denote the ratio of informed to uninformed consumers.

At the beginning of the period, the firm and informed consumers observe the true quality of the product. The firm then sets a price P , and uninformed consumers update their beliefs about product quality on the basis of this signal. Let $b = b(P)$ be the uninformed consumers' posterior belief that quality is high when the price is P . Consumers are assumed to make purchase decisions that maximize expected utility (i.e., the expected reservation price minus P), given beliefs. This process generates an informed demand curve, in which a fraction $1 + P^L - P$ of informed consumers buy when $P \in [P^L, 1 + P^L]$ and $q = H$, and an uninformed demand curve, characterized by a fraction $1 + (P^L - P)/b$ of uninformed consumers buying when $P \in [P^L, b + P^L]$ and quality is believed to be high with probability b . With these demand curves and our assumptions on cost technologies, the profit of a firm with quality q and price P facing uninformed consumers with belief b , denoted $\pi(q, b, P)$, is straightforward to define explicitly.⁴ We assume that the objective of the firm is to maximize profits.

³Also, Jean Tirole (1988 Ch. 2) has a nice expository discussion of price as a signal of quality which refers to an earlier working paper (Bagwell and Riordan, 1986), which this article supersedes.

⁴For example, if $P \in (P^L, b + P^L)$, then $\pi(L, b, P) = P[1 + (P^L - P)/b]M/(1 + X)$ and $\pi(H, b, P) = (P - c)[(1 + P^L - P)X + (1 + (P^L - P)/b)]M/(1 + X)$.

These actions and objectives define an extensive-form game of incomplete information with multiple sequential equilibria (Kreps and Robert Wilson, 1982). A sequential equilibrium requires that the firm and consumers act in a sequentially rational fashion and that uninformed consumers update beliefs using Bayes' rule on the equilibrium path. As usual, we distinguish between separating equilibria (in which high- and low-quality firms choose different prices) and pooling equilibria. However, we do restrict attention to pure-strategy equilibria.

We select plausible equilibria by imposing the "intuitive criterion" (Cho and Kreps, 1987). Consider an equilibrium in which the firm earns profits of $\pi(H)$ and $\pi(L)$ for high- and low-quality products, respectively. Then the equilibrium satisfies the intuitive criterion if there does not exist a price P' such that: (a) $\pi(H, 1, P') > \pi(H)$ and (b) $\pi(L, 1, P') < \pi(L)$. Intuitively, if such a price P' did exist, then uninformed consumers should believe that only a high-quality firm would charge P' , which by (a) causes the equilibrium to fail.⁵

Letting $P(q)$ denote the equilibrium price charged by a type- q firm, we now state our first two lemmas, which are almost obvious.

LEMMA 1: *In any equilibrium, $P(q) \geq P^L$ for $q \in \{H, L\}$.*

PROOF:

The function $\pi(q, b(P), P)$ is strictly increasing in P , for all $q \in \{H, L\}$ and all functions $b(\cdot)$, when $P < P^L$. Thus, were $P(q)$ less than P^L , the type- q firm could increase its price slightly and increase profits.

LEMMA 2: *In any separating equilibrium, $P(H) > P^L$ and $P(L) = P^L$.*

⁵We use the intuitive criterion throughout to restrict the class of equilibria. A weaker refinement is to eliminate dominated strategies (Elon Kohlberg and Jean-Francois Mertens, 1986; Milgrom and Roberts, 1986). This refinement is sufficient for all of our results for separating equilibria and for some of our work on pooling equilibria. The intuitive criterion is necessary, however, for Theorems 3 and 6.

PROOF:

A low-quality firm earns zero profits in a separating equilibrium if $P(L) > P^L$ and positive profits if $P(L) = P^L$. Therefore, the results follows from Lemma 1.

In a separating equilibrium, the low-quality firm charges P^L and the high-quality firm charges some higher price. Moreover, it must be that $\pi(L, 0, P^L) \geq \pi(L, 1, P(H))$; otherwise, the low-quality firm would mimic its high-quality counterpart. We are thus led to consider the set

$$\{P | \pi(L, 0, P^L) = \pi(L, 1, P)\}.$$

This equation has an upper and lower root,

$$\bar{P}(X) = [(1 + P^L)/2]$$

$$+ \left(\left[(1 + P^L)^2/4 \right] - P^L(1 + X) \right)^{1/2}$$

and

$$\underline{P}(X) = [(1 + P^L)/2]$$

$$- \left(\left[(1 + P^L)^2/4 \right] - P^L(1 + X) \right)^{1/2}$$

expressed as functions of X .

We assume that $1 > P^L$ and represent $\bar{P}(X)$ and $\underline{P}(X)$ by the upper and lower boundaries of the parabola in Figure 1.⁶ These equations have no solution for values of $X > \bar{X}$. For $X \leq \bar{X}$, any price inside the parabola, $P \in (\underline{P}(X), \bar{P}(X))$, satisfies $\pi(L, 0, P^L) < \pi(L, 1, P)$; the low-quality firm would mimic any such price. This leads immediately to the following lemma.

⁶If $P^L \geq 1$, then for $\varepsilon = P - P^L > 0$, $\pi(L, 1, P) = \pi(L, 0, P^L) - [P^L X + \varepsilon(P^L - 1) + \varepsilon^2]M/(1 + X) < \pi(L, 0, P^L)$, so that a low-quality firm's marginal revenue for a price increase above P^L is negative no matter how beliefs adjust. Thus, if $P^L \geq 1$, then mimicry is never profitable for the low-quality firm, and the problem degenerates. The following can be shown: if $P^L \geq 1 + c > 1$, then the unique equilibrium has $P(H) = P(L) = P^L$; while if $1 + c \geq P^L \geq 1$, then the unique equilibrium has $P(L) = P^L$ and $P(H) = (1 + P^L + c)/2 \equiv P^H$. $P^L < 1$ is thus the interesting case.

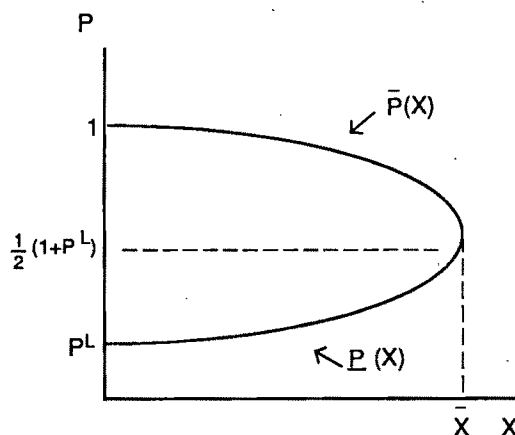


FIGURE 1. SEPARATING PRICES

LEMMA 3: *If $X < \bar{X}$, then in any separating equilibrium either $P(H) \geq \bar{P}(X)$ or $P(H) \leq \underline{P}(X)$.*

It is important to understand the parabola in Figure 1. By mimicking a high-quality price $P(H) > P^L$, the low-quality firm both gains and loses. It gains by tricking uninformed consumers with reservation prices at or above $P(H)$ into buying at a higher price, but it loses because all informed consumers and the remaining uninformed consumers refuse to buy at that price. The gains outweigh the losses for prices inside the parabola; for these prices, the low-quality firm finds mimicry profitable.

When all consumers are uninformed ($X = 0$), a price slightly above P^L is certainly worth mimicking, as only a very few consumers refuse to buy at the higher price. It follows that the lower branch of the parabola begins at P^L , reflecting the fact that all uninformed consumers buy at this price regardless of beliefs. As price climbs higher, eventually sales to uninformed consumers are sufficiently restricted that mimicry becomes unattractive. The critical price marking the beginning of the upper branch of the parabola is easily shown to be 1. When the ratio of informed to uninformed consumers (X) rises, it becomes increasingly costly for the low-quality firm to mimic its

high-quality counterpart and thereby sacrifice all informed purchases. For this reason, the parabola narrows about $(1 + P^L)/2$, the maximizer of $\pi(L, 1, P)$. This is the low-quality firm's monopoly price when all uninformed consumers believe it to be high-quality. For X above \bar{X} , the low-quality firm refuses to mimic even $(1 + P^L)/2$.

The high-quality firm's full-information monopoly price is $P^H \equiv (1 + P^L + c)/2$, the maximizer of $\pi(H, 1, P)$. However, with uninformed consumers, the high-quality price may be distorted upward as shown in the following theorem. The theorem establishes necessary conditions for a separating equilibrium satisfying the intuitive criterion.

THEOREM 1: $P(H) = \max\{\bar{P}(X), P^H\}$ and $P(L) = P^L$ are the only separating equilibrium prices satisfying the intuitive criterion.

PROOF:

Lemma 2 implies $P(L) = P^L$, so suppose $P(H) \neq \max\{\bar{P}(X), P^H\}$ and consider Figure 2. X^H satisfies $\bar{P}(X^H) = P^H$. For $X > X^H$, the intuitive criterion fails by setting $P' = P^H$. For $X \leq X^H$, Lemma 3 rules out $P(H) \in (\underline{P}(X), \bar{P}(X))$. Moreover, the intuitive criterion fails for $P' \in (\bar{P}(X), P(H))$ if $P(H) > \bar{P}(X)$, and for $P' \in (P(H), \underline{P}(X))$ if $P(H) < \underline{P}(X)$. This leaves the possibility that $P(H) = \underline{P}(X)$, but it is straightforward to show that $c > 0$ implies $\pi(H, 1, \bar{P}(X)) > \pi(H, 1, \underline{P}(X))$, from which it follows that $P' = \bar{P}(X) + \varepsilon$ violates the intuitive criterion for sufficiently small ε .

Theorem 1 identifies two cases. If $P^H \geq \bar{P}(X)$, then the separating equilibrium has $P(H) = P^H$ and $P(L) = P^L$. This is because the low-quality firm is unwilling to mimic the high-quality firm's favorite price. The more interesting case emerges when $P^H < 1$, or equivalently $P^L + c < 1$, and separation is potentially costly for the high-quality firm. In this case, a supramonopoly price is charged if the ratio of informed to uninformed consumers is small.

COROLLARY 1: *If $P^H < 1$ and X is sufficiently small, $P(H) = \bar{P}(X) > P^H$ and $P(L) = P^L$ are the only separating equilibrium prices satisfying the intuitive criterion.*

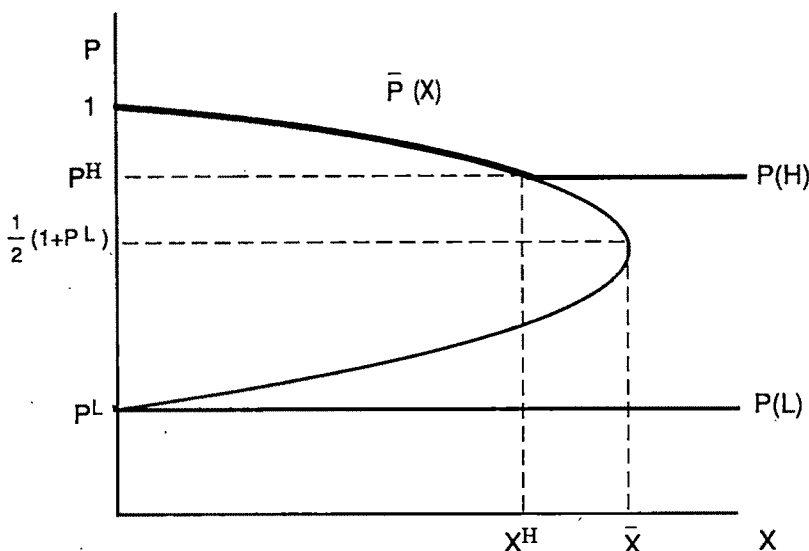


FIGURE 2. EQUILIBRIUM SEPARATING PRICES

For expositional reasons, we assume henceforth that $P^H < 1$.

Figure 2 illustrates the prices charged in a separating equilibrium satisfying the intuitive criterion (the parabola is the same as in Fig. 1). It may seem surprising that a high-quality firm separates with $\bar{P}(X)$ instead of a lower price $\underline{P}(X)$, when X is small. However, a simple intuition underlies this result. Because high-quality production is costly ($c > 0$), the full-information monopoly price P^H is closer to $\bar{P}(X)$ than to $\underline{P}(X)$.⁷ Put differently, the high price is the efficient means of separation because the forgone profit from a lost customer is less for the high-quality firm (Milgrom and Roberts, 1986). Thus, the high-quality firm prefers to separate with the high price.

⁷The assumption that higher-quality firms have at least somewhat higher costs of production is crucial. As the cost asymmetry between qualities diminishes, the high-quality firm's preferred separating price, $\bar{P}(X)$, remains constant; that is, $\bar{P}(X)$ is independent of c . In the limiting case when $c = 0$, prices corresponding to the lower branch of the parabola also satisfy the intuitive criterion. This is because the high-quality firm is then indifferent between $\bar{P}(X)$ and $\underline{P}(X)$. We believe $c > 0$ to be a realistic case.

So far we have characterized necessary conditions for a separating equilibrium. We now turn to existence.

Separation can occur only if the high-quality firm chooses not to monopolize informed consumers at the expense of losing uninformed consumers. Such a deviation is potentially attractive only if separation is costly for the high-quality firm, that is, when $X < X^H$ (see Fig. 2) in which case $P(H) = \bar{P}(X) > P^H$. Prices that might then increase high-quality profit must be inside the parabola and thus must also be prices that could increase low-quality profit. The intuitive criterion does not restrict beliefs for such prices, and at worst a deviation in this range could induce the belief of certain low quality. The high-quality firm will thus charge the price $P(H) = \bar{P}(X)$ when $X < X^H$ in a separating equilibrium if and only if $\pi(H, 1, \bar{P}(X)) \geq \pi(H, 0, P^H)$.⁸ Setting $\pi(H, 1, P) = \pi(H, 0, P^H)$ when $X < X^H$ de-

⁸ P^H is clearly the best price above P^L for the high-quality firm to deviate to when uninformed consumers believe quality is low. The high-quality firm will never deviate from $\bar{P}(X)$ to P^L (or lower), since $\pi(H, 1, \bar{P}(X)) \geq \pi(H, 1, \bar{P}(0)) > \pi(H, 1, P(0)) = \pi(H, 1, P^L)$.

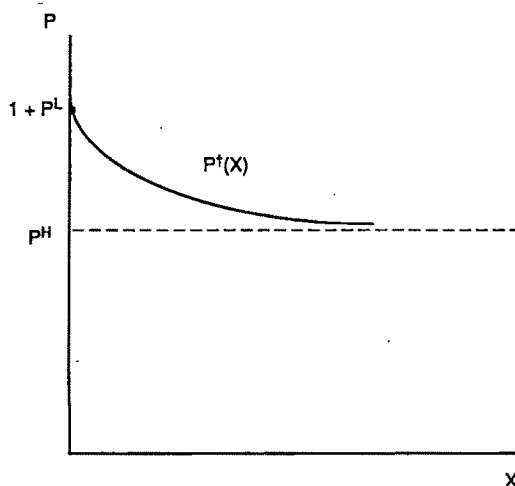


FIGURE 3. NO-DEFECT PRICES

finds a "no-defect" root,

$$P^+(X) = P^H + [(1 + P^L - c)(1 + X)^{-1/2}/2]$$

which begins at $(1 + P^L)$ for $X=0$ and asymptotically declines to P^H , as shown in Figure 3. The high-quality firm has no incentive to defect if and only if $P(H) \leq P^+(X)$. Since "intuitive" beliefs can entail $b(P')=0$ for all $P' \in (P(X), \bar{P}(X))$, it is easily established that the low-quality firm is also unwilling to defect from the proposed separating equilibrium. We thus have the following existence theorem.

THEOREM 2: *A separating equilibrium satisfying the intuitive criterion exists if and only if $X \geq \bar{X}$ or if $X < \bar{X}$ and*

$$\Delta(X) \equiv P^+(X) - \bar{P}(X) \geq 0.$$

Two observations are in order, which we state below as corollaries. First, a separating equilibrium always exists if the ratio of informed to uninformed consumers is small because $\Delta(0) \equiv P^L > 0$. Second, it can be shown numerically that a separating equilibrium exists for any value of X unless P^L and c are small; Figure 4 illustrates parameter values for which separating equilibria fail to exist for some intermediate X .

COROLLARY 2: *A separating equilibrium satisfying the intuitive criterion exists if X is sufficiently small.*

COROLLARY 3: *A separating equilibrium satisfying the intuitive criterion exists if P^L or c is sufficiently large.*

We next turn to pooling equilibria. The following theorem establishes that, if the percentage of informed consumers is sufficiently small, then the only equilibrium satisfying the intuitive criterion is a separating equilibrium. In other words, when the market is very uninformed, there always exists a high price at which the high-quality firm can profitably distinguish itself.

THEOREM 3: *If X is sufficiently small, then no pooling equilibrium satisfies the intuitive criterion.*

PROOF:

We prove the result for $X=0$; the result follows for X close to 0 by continuity. Let $Q(P, b) \equiv [1 - (P - P^L)/b]M$ denote the quantity of sales at a price P when consumers believe high quality with probability b . In a pooling equilibrium, $P(H) = P(L) = P^*$, and $b(P^*) = r$. The high- and low-quality firms earn profits $\pi(H) \equiv (P^* - c)Q(P^*, r)$ and $\pi(L) \equiv P^*Q(P^*, r)$, respectively. Clearly $1 + r > P^* \geq \max\{P^L, c\}$; otherwise, one or the other type of firm would defect. Thus, $\pi(L) > \pi(H) \geq 0$. Since $P^*Q(P^*, 1) > \pi(L)$, there exists $P'' > P^*$ such that $\pi(L, 1, P'') \equiv P''Q(P'', 1) = \pi(L)$ and $\pi(H, 1, P'') = \pi(H) + c[Q(P^*, r) - Q(P'', 1)] > \pi(H)$. Moreover, $PQ(P, 1)$ is decreasing in P at P'' . Therefore, $P' \equiv P'' + \varepsilon$ violates the intuitive criterion for $\varepsilon > 0$ sufficiently small.

Pooling is therefore impossible if X is sufficiently small. Similarly, if the market is sufficiently well informed that X is large, then pooling is impossible to maintain, as each firm type prefers to deviate and monopolize informed consumers. For example, if $X \geq \bar{X}$ (see Fig. 1), then pooling is clearly impossible, because the low-quality firm would not select $P > P^L$ even if it were

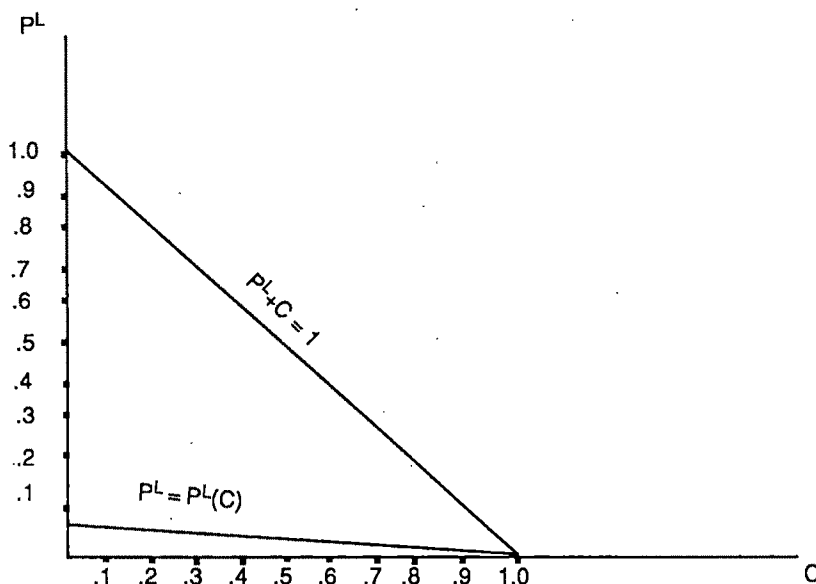


FIGURE 4. PARAMETER VALUES SUPPORTING SEPARATION: FOR $\{P^L, C\}$ SUCH THAT $P^L > P^L(C)$, A UNIQUE SEPARATING EQUILIBRIUM SATISFYING THE INTUITIVE CRITERION EXISTS; IF INSTEAD $P^L < P^L(C)$, A SEPARATING EQUILIBRIUM MAY FAIL TO EXIST FOR INTERMEDIATE VALUES OF X

then believed to certainly have a high-quality product. The critical X beyond which pooling is impossible is actually smaller than \bar{X} , since pooling only generates the belief $b(P(L)) = r$. This point is stated in the following theorem, as is the related point that pooling is impossible when consumers' prior belief of high quality is pessimistic and, correspondingly, the profits from pooling are low.⁹

THEOREM 4: *If $r \leq \max\{P^L, c - P^L\}$ or $X \geq (r - P^L)^2 / 4rP^L$, then no pooling equilibrium exists satisfying the intuitive criterion.*

⁹By Theorems 2, 3, and 4 and Corollary 2, pure-strategy sequential equilibria satisfying the intuitive criterion might not exist if r is small and X is in an intermediate range. However, equilibria do exist satisfying the intuitive criterion in which the high-quality firm selects a price $P(H) > P^L$, the low-quality firm mixes between $P(H)$ and P^L with weights λ and $1 - \lambda$, and uninformed consumers believe high quality with probability b , where $P(H)$, λ , and b satisfy $b = r/[r + (1 - r)\lambda]$, $\pi(H, b, P(H)) = \pi(H, 0, P^H)$, and $\pi(L, b, P(H)) = \pi(L, 0, P^L)$.

PROOF:

By Lemma 1, pooling can never occur at $P < P^L$. Moreover, since $1 > P^L$, pooling at $P = P^L$ violates the intuitive criterion as the high-quality firm would deviate to $\bar{P}(X)$. A necessary condition for pooling at P is $\pi(L, r, P) \geq \pi(L, 0, P^L)$. Suppose pooling occurs at $P > P^L \geq r$. Let $M/(1 + X)$ denote the stock of uninformed consumers and let $\varepsilon \equiv P - P^L > 0$. Then $\pi(L, r, P) = \pi(L, 0, P^L) - [\varepsilon(P^L - r)/r + \varepsilon^2/r]M/(1 + X) < \pi(L, 0, P^L)$, a contradiction. Suppose next that pooling occurs when $r \leq c - P^L$. Pooling at P requires $P \geq c$ and $P < r + P^L$ (lest the low-quality firm deviate to P^L), which is contradictory. Finally, suppose $r > \max\{P^L, c - P^L\}$. Then it is easy to show that $X \geq (r - P^L)^2 / 4rP^L$ implies that at $P > P^L$, $\pi(L, r, P) \leq \pi(L, r, (r + P^L)/2) < \pi(L, 0, P^L)$, a contradiction.

Finally, if r is big and X is intermediate, pooling equilibria satisfying the intuitive criterion may exist. The possible prices for such equilibria are easily restricted. First, pooling at or below P^L is inconsistent with the intuitive criterion, as argued in the proof

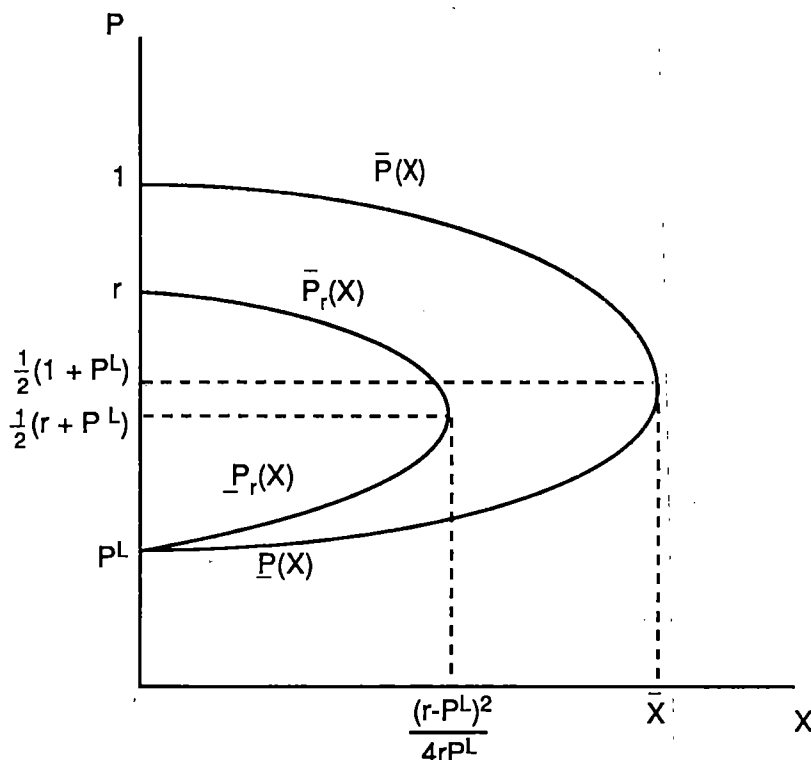


FIGURE 5. THE POOLING PARABOLA

of Theorem 4. Second, given that pooling must occur above P^L , the pooling price must certainly be inside the parabola, or the low-quality firm would deviate to P^L . In fact, since pooling only gives the belief $b(P(L)) = r$, a tighter bound can be found. Setting $\pi(L, r, P) = \pi(L, 0, P^L)$ gives the prices at which the low-quality firm is just willing to pool. The corresponding roots are

$$\bar{P}_r(X) = [(r + P^L)/2]$$

$$+ \left(\left[(r + P^L)^2/4 \right] - rP^L(1 + X) \right)^{1/2}$$

$$\underline{P}_r(X) = [(r + P^L)/2]$$

$$- \left(\left[(r + P^L)^2/4 \right] - rP^L(1 + X) \right)^{1/2}$$

As shown in Figure 5, a "pooling parabola" inside the initial (separating) parabola is thus defined. The pooling parabola is drawn under the assumption that $r > P^L$ and is not defined for $X \geq (r - P^L)^2/4rP^L$, as suggested by Theorem 4. Any price outside of the pooling parabola has $\pi(L, r, P) < \pi(L, 0, P^L)$ and therefore cannot be supported as a pooling equilibrium. We thus have the following necessary condition for pooling.

THEOREM 5: *In any pooling equilibrium satisfying the intuitive criterion, $\underline{P}_r(X) \leq P(H) = P(L) \leq \bar{P}_r(X)$.*

Therefore, if an intuitive pooling equilibrium does exist, then r must be large, X must be intermediate, and the pooling price must be lower than the intuitive separating price for high quality.

II. Multiperiod Extensions

A. Interpreting the Static Results

We begin by offering a multiperiod interpretation of the static results of the previous section. Suppose that the market evolves in two periods. In the introductory period, the market is poorly informed about the characteristics of the product; the ratio of informed to uninformed consumers (X_1) is close to zero. Over time, information about the product diffuses through published quality reviews (e.g., *Consumer Reports*). Thus, during the mature phase of the market, the ratio of informed to uninformed consumers (X_2) is larger.

A high-quality product ($q = H$) will be introduced at a price that is distorted above the full-information monopoly price (Corollary 1 and Theorem 3). As information diffuses, it becomes easier to signal high quality, and the size of this distortion is reduced. If X_2 is sufficiently large, then a lower price signals high quality during the mature period (Theorem 4). For intermediate values of X_2 , pooling might occur during the mature phase; however, even a pooling price must be below the introductory price (Theorem 5). Therefore, high-quality products exhibit high and declining prices in either case.

A low-quality product, on the other hand, is introduced and remains at a low price P^L if X_1 is small and X_2 is large (Theorems 1, 3, and 4). If, instead, X_1 is small and X_2 is in an intermediate range, then the product is introduced at a low price but might rise to a higher pooling price in the mature phase (Theorems 1 and 5). However, this possibility is unlikely to be observed, because pooling can occur only if the prior probability of low quality is small (Theorem 4). Therefore, we conclude that, on average, prices will tend to fall as the market evolves.

In summary, our results appear to be consistent with the stylized facts for consumer durables (see the Introduction). Consider a reasonably well-defined product category, such as microwave ovens or color TVs. Over time, quality improvements lead

to the introduction of new products, but information about the value of these improvements diffuses gradually. New, higher-quality products are introduced at higher prices, resulting in a positive rank-order correlation between price and quality. As consumers gain information about new products, prices on average decline over time, since it becomes easier or unnecessary to signal quality, and price differentials narrow.

Finally, a weakened correlation between price and quality might be explained by measurement errors in the data. To understand this point, suppose that price and quality were perfectly rank-order correlated but that price differentials narrowed over time. Suppose further that prices and quality were measured with error and that this measurement error was independently and identically distributed over time. Then the measured rank-order correlation between price and quality would tend to decline.

This interpretation of our static results can be formalized in a two-period extensive-form game under the following assumptions. First, the populations of consumers in each of the two periods are distinct. That is, a consumer enters the market at the beginning of one or the other period, either makes a purchase or does not, and then leaves the market at the end of that period. In other words, we are taking a long-run view of the evolution of the market.

Second, upon entering the market, consumers only observe the current price. Specifically, consumers in the mature phase do not know what price was charged by the firm in the introductory phase. They do, however, know the date, that is, whether they are in the introductory or mature phase. (We relax these assumptions below.) As new consumers are ignorant of past prices, we refer to this model as the *ignorant-consumer model*.

Finally, information diffusion is independent of market activity. During the introductory phase, most consumers cannot distinguish a high-quality product prior to consumption, although some small fraction might be able to do so, say through inspec-

tion, because of some related expertise. During the mature phase, a new population has better information about product quality, because of the availability of published quality reviews. However, the ratio of informed to uninformed consumers in the mature period is independent of the quantity of sales in the first period.

Under these assumptions, the multi-period model decomposes into a sequence of one-period models, with our results of the previous section applying to each period. The fundamental dynamic variable is the ratio of informed to uninformed consumers, X_t , which by assumption increases over time (Bagwell and Riordan, 1986).

The assumption that information diffusion is independent of market activity might be important for our basic conclusion that high and declining prices signal product quality. Suppose instead that the number of informed consumers during the mature phase depends on the number of introductory sales. This might be the case if word-of-mouth communication about personal experience were an important mechanism for information diffusion. In this case, high prices which discourage sales become a less attractive method of signaling high quality. The high-quality firm might instead prefer to signal with a low price, in which case price would tend to rise over time.

However, word-of-mouth communication is not a particularly good method of information diffusion for many products. For example, for many durable products (e.g., smoke alarms), it is difficult to determine quality even after purchase. For such products, quality reviews or the advice of experts are much more important sources of information diffusion.

Further, word-of-mouth communication may be a very noisy source of information if there is an idiosyncratic component to consumer tastes. For example, knowing that my neighbor likes his new microwave oven may not be a very good indicator that I will like it. On the other hand, knowing that 90 percent of consumers like it is much more informative. Published quality reviews, such as those in *Consumer Reports*, accomplish this information aggregation. Our model

could be extended to introduce an idiosyncratic taste component (Milgrom and Roberts, 1986).¹⁰

We conclude that our independence assumption is consistent with a long-run view of markets, in which the primary source of information diffusion is published quality reviews. This seems particularly pertinent for explaining stylized facts based on data published in *Consumer Reports*. It is plausible that the availability of published quality reviews is independent of first-period pricing.

We turn now to a relaxation of some other assumptions about consumer information.

B. Confused Consumers

In the above-described ignorant-consumer model, the ratio of informed to uninformed consumers is common knowledge. In our model, this is equivalent to assuming that each consumer knows the age of the firm, since by assumption X_t increases strictly over time. This is clearly an extreme assumption, as it seems reasonable that consumers might also be incompletely informed about the demand side of the market. We address this issue in a very simple way by assuming that consumers in any period observe only the current price and not the age of the firm. We continue to assume that some fraction of consumers are informed about quality in each period. As consumers are confused about the date, we refer to this model as the *confused-consumer model*.

An important novelty of the confused-consumer model is that the firm now has two dimensions of private information: the

¹⁰The assumption that initial sales volume influences the future information state is in fact problematic, even if quality is nonidiosyncratic and easily evaluated after experience. In particular, in a separating equilibrium, each current consumer believes he knows quality whether or not he chooses to buy. Word-of-mouth communication then affects the rate of diffusion only if a consumer who knows quality communicates more effectively than does a consumer who believes himself to know quality.

quality of its product and the prevailing ratio of informed to uninformed consumers. Consequently, the price set by the firm potentially signals information about both. Uninformed consumers will thus attempt to update their beliefs about both unknown variables.

Formally, we model this situation as a two-period extensive-form game with the following structure. At any period t , the ratio of informed to uninformed consumers is actually X_t . The firm knows this ratio as well as the quality of its product and chooses a price, $P_t(q)$. Uninformed consumers at any date do not observe quality and also do not know whether they are in period one or two. That is, an uninformed consumer knows that $q \in \{H, L\}$ and $X \in \{X_1, X_2\}$, where H , L , X_1 and X_2 are all commonly known values and $X_2 > X_1$, but he does not know the actual realizations of q and X . Upon observing a price P , an uninformed consumer forms a (stationary) belief $b(P)$, representing the likelihood of high quality. In a sequential equilibrium, the belief of uninformed consumers must agree with Bayes' rule for prices that could occur (for some quality and date) in equilibrium.¹¹

The intuitive criterion can be usefully employed in this environment. A sequential equilibrium fails the intuitive criterion if there exists an out-of-equilibrium price P' such that (a) $\pi_t(H, 1, P') > \pi_t(H)$ for some t and (b) $\pi_t(L, 1, P') < \pi_t(L)$ for all t , where $\pi_t(q)$ are equilibrium profits at date t for quality q , with $t \in \{1, 2\}$ and $q \in \{H, L\}$. Intuitively, if there exists an out-of-equilibrium price that is profitable for a high-quality firm at some date when uninformed

consumers believe it to be high-quality, but which is never profitable for a low-quality firm no matter what uninformed consumers believe, then uninformed consumers must believe that there is high quality at that price. In effect, the firm makes an implicit speech as to quality and date with the selection of the price P' .

We will continue to interpret the first period as an introductory phase in which X_1 is close to zero and the second period as a mature phase with X_2 large. Our main result is that the high-quality firm will separate with a high price in period one, and the prices of both types of firm will decline. This equilibrium exhibits intertemporal pooling, in that the young, low-quality firm mimics the price of the mature, high-quality firm.¹² Thus, the general implications of the ignorant-consumer model also are true in the confused-consumer model.

THEOREM 6: *In the confused-consumer model, if X_1 is sufficiently small, then in any equilibrium satisfying the intuitive criterion, $P_1(H) \geq P_2(H)$, $P_1(H) \in (P^H, \bar{P}(X_1)]$, $b(P_1(H)) = 1$, and $P_1(L) \geq P_2(L)$. If in addition X_2 is sufficiently large, then $P_1(H) > P_2(H)$ and $P_1(L) = P_2(H) > P_2(L) = P^L$.*

The theorem has a simple intuition. (A formal proof is in the Appendix.) The declining path of prices follows from the assumption that the number of informed consumers increases through time. As the market becomes more informed, distortions away from complete-information monopoly prices become more costly. Thus, a rising price profile is not an optimal strategy for a firm, since it entails greater distortions when the market is more informed. The firm would be better off, for example, to reverse the order of its two prices. The assumption that the number of informed consumers is initially small is not important in ruling out rising price profiles but does ensure that the introductory high-quality price separates; the logic is similar to that for Theorem 3.

¹¹The specific nature of the consumer arrival process will determine the consumers' priors over dates. It is perhaps easiest to imagine that the process treats consumers anonymously, pulling out a certain fraction in period one and the remainder in period two, in which case consumers have common priors over dates (Bagwell and Riordan, 1986). The consumers' priors over dates will affect beliefs as to quality only for prices that would be charged at different dates by the high- and low-quality firms. In any event, Theorem 6 below characterizes the necessary characteristics of equilibrium and is independent of the details of the arrival process.

¹²Bagwell and Riordan (1986) give an example of an equilibrium with these characteristics.

It is interesting to note that the initial high-quality price in the confused-consumer model is never higher and, indeed, may be lower than the corresponding price in the ignorant-consumer model when X_1 is small. Since the introductory period is when the low-quality firm has the greatest incentive to mimic, the price $\bar{P}(X_1)$ is sufficient to separate even when consumers are confused as to the date. Furthermore, an introductory price strictly below $\bar{P}(X_1)$ may separate when consumers are confused, because the low-quality firm initially earns a larger profit by mimicking the mature, high-quality price than by charging P^L .

C. Hindsightful Consumers

We close this section by briefly considering another assumption, that consumers observe both the age of the firm and the entire history of prices; we call this model the *hindsightful-consumer model*. This is a strong assumption (e.g., past prices are not typically reported in *Consumer Reports*).

In this model, uninformed consumers in period one observe a first-period price P_1 and attempt to infer quality, while uninformed consumers in period two base their beliefs on a pair of prices, P_1 and P_2 . Sequential equilibrium then requires that the beliefs of an uninformed consumer be consistent with Bayes' rule for any price or pair of prices that could occur in equilibrium. The sequential equilibria of the ignorant-consumer model are also equilibria of this model but have the property that consumers always ignore past prices in forming beliefs about product quality. However, there also exist other equilibria in which period-two consumers base their quality beliefs at least partially on period-one price.

Cho (1987) has generalized the intuitive criterion to multiperiod settings such as this. However, as Cho (1987 p. 1385) notes, this refinement is often very weak, and this is true in our model as well.¹³ Nevertheless,

there do exist plausible sequential equilibria—other than those of the ignorant consumer model—that are consistent with the stylized conclusion that high and declining prices signal quality.¹⁴

A particularly focal equilibrium is that which is the Pareto dominant separating equilibrium for the firm. To find this equilibrium, we solve

$$\max_{(P_1, P_2)} [\pi_1(H, 1, P_1) + \delta \pi_2(H, 1, P_2)]$$

subject to

$$\pi_1(L, 1, P_1) \leq \pi_1(L, 0, P^L)$$

and

$$\begin{aligned} \pi_1(L, 1, P_1) + \delta \pi_1(L, 1, P_2) \\ \leq \pi_1(L, 0, P^L) + \delta \pi_2(L, 0, P^L) \end{aligned}$$

where δ is a common discount factor. Our fundamental result is that the Pareto dominant separating equilibrium is characterized by high and declining high-quality prices, while low-quality prices are constant at P^L . The price dynamics for the hindsightful-consumer model are again qualitatively the same as for the ignorant-consumer model. In each case, the high-quality price begins high and then declines as the number of informed consumers increases and signaling becomes easier.

THEOREM 7: *In the hindsightful-consumer model, the Pareto dominant separating equilibrium is characterized by $P_1(H) \geq \bar{P}(X_1)$, $P_2(H) = P^H$ if $X_2 \geq X^H$, $P_2(H) \in [\bar{P}^H, \bar{P}(X_2)]$ if $X_2 < X^H$, and $P_1(L) = P_2(L) = P^L$.*

consumers. It follows that, using the intuitive criterion, period-one consumers are generally unable to associate a deviation with the high-quality product.

¹⁴For a more thorough discussion of the types of plausible separating equilibria in the hindsightful-consumer model, see Bagwell and Riordan (1989).

¹³The problem is that almost any deviant first-period price could improve profits for the low-quality firm, if the deviation were favorably interpreted by period-two

The proof of Theorem 7 is straightforward (Bagwell and Riordan, 1989).

It is interesting to compare the Pareto dominant separating equilibrium for the hindsight-consumer model with the separating path in the ignorant-consumer model. Numerical calculations reveal that the two paths agree if P^L is large. However, if P^L is small and X_2 is not too large, the hindsight-consumer model's equilibrium has $P_1(H) > \bar{P}(X_1)$ and $P_2(H) < \bar{P}(X_2)$. Here, the high-quality firm "front loads" the signaling process by signaling strongly in the introductory phase and then profit-taking in the mature phase.¹⁵

A general theme seems to arise from a comparison of the ignorant-, confused-, and hindsight-consumer models. In the focal separating equilibria of these models, the initial high-quality price is at least as high in the hindsight- as in the ignorant-consumer model, and is at least as high in the ignorant- as in the confused-consumer model. In effect, the more information consumers have about the initial period, the more difficult and the more important it is for the high-quality firm to signal strongly in that period. Since signaling involves high prices, the initial high-quality price is highest when consumers have hindsight and lowest when they are confused.

III. Conclusions

A high-quality good will be introduced at a high price that is lowered over time toward the full-information monopoly price. The high introductory price signals high quality, because a high-cost firm is more

willing to restrict sales volume than is a low-cost firm. Furthermore, a low-quality firm loses greater sales volume from a high price, since informed consumers refuse to buy at such a price. As information about product quality diffuses and more consumers become more informed, it therefore becomes easier for a high-quality firm to signal its quality. High-quality prices thus decline as the market matures.

Our prediction of high and declining prices is consistent with the stylized facts of the marketing literature.¹⁶ In particular, our model provides an explanation for the declining trends in consumer durables of real prices, price differentials between competing brands, and the rank-order correlation between price and quality.

The prediction is robust to a variety of assumptions about consumer information. Whether or not uninformed consumers know past prices and firm age, intuitively plausible equilibria exist in which high-quality products have high and declining prices.

Our model relies on many special features that can be relaxed. For example, linear demands and costs are not crucial for our conclusions.¹⁷ The two-period framework is also easily extended into a many-period setting.

Many interesting extensions do remain. It would be worthwhile to study an explicit model of word-of-mouth communication. Further empirical work might investigate how price-quality relationships depend on the amount and type of consumer information. One intriguing possibility is to interpret the publication of a *Consumer Reports* evaluation of a product group as a proxy for an increase in the "fraction of informed consumers." This suggests the empirical test that a *Consumer Reports* publication should itself lower the relative price of high-quality items and weaken price-quality correlation.

¹⁵More generally, since the ignorant-consumer equilibrium is not always the hindsight-consumer equilibrium, the high-quality firm has a clear incentive to ensure that past price information is available to uninformed consumers. By equipping consumers with hindsight, the high-quality firm may be able to generate a more profitable equilibrium. This point is reminiscent of work by B. Douglas Bernheim and Michael D. Whinston (1990), who argue that firms may seek multi-market contact so as to pool incentive constraints. Analogously, we find that a high-quality firm may want consumers to observe its past pricing decisions so as to pool its "no-mimic" constraints through time.

¹⁶It should be noted, however, that at least some of our "stylized facts" are in fact controversial (Valerie A. Zeithaml, 1988).

¹⁷Bagwell (1990) establishes that high prices signal a high-quality product in a model that allows for a general demand function and product quality choice.

APPENDIX

PROOF OF THEOREM 6:

Observe first that $P_t(H) > P^L$ and $P_t(L) \geq P^L$. $P_t(q) < P^L$ is impossible in equilibrium, as $P_t(q) = P^L$ earns greater profit always. Also, $P_t(H) = P^L$ is inconsistent with the intuitive criterion, because $\pi_t(H, 1, \bar{P}(X_1)) - \pi_t(H, b(P^L), P^L) > 0$ and $\pi_t(L, 1, \bar{P}(X_1)) \leq \pi_t(L, b(P^L), P^L) \leq \pi_t(L)$.

Next, X_1 small implies $b(P_1(H)) = 1$. To see this, suppose $b(P_1(H)) < 1$. From the proof of Theorem 3, there exists $P'' > P_1(H)$ such that $\pi_t(L, 1, P'') = \pi_t(L, b(P_1(H)), P_1(H)) \leq \pi_t(L)$ and $\pi_1(H) = \pi_1(H, b(P_1(H)), P_1(H)) < \pi_1(H, 1, P'')$. $P' = P'' + \varepsilon$ with $\varepsilon > 0$ and small, then violates the intuitive criterion.

Given $b(P_1(H)) = 1$, $P_1(H) \in (P^H, \bar{P}(X_1)]$ follows. To see this, note that $P_1(H) = P^H$ induces $P_2(H) = P^H$ and thus $P_1(L) = P^H$, which contradicts $b(P_1(H)) = 1$. Further, $P_1(H) < P^H$ violates the intuitive criterion; for if $P_1(H) \in [(1 + P^L)/2, P^H]$, then $P' = P_1(H) + \varepsilon$, for $\varepsilon > 0$ and small, violates the intuitive criterion. On the other hand, if $P_1(H) < (1 + P^L)/2$, then as in the proof of Theorem 1, $P'' > P^H$ can be found for which $\pi_t(L, 1, P'') = \pi_t(L, 1, P_1(H)) \leq \pi_t(L)$ and $\pi_1(H, 1, P'') > \pi_1(H, 1, P_1(H)) = \pi_1(H)$. $P' = P'' + \varepsilon$ then violates the intuitive criterion. Finally, $P_1(H) > \bar{P}(X_1)$ is also inconsistent with the intuitive criterion, as can be seen by setting $P' = P_1(H) - \varepsilon$.

$P_1(H) \geq P_2(H)$ is straightforward to establish. Otherwise, since $b(P_1(H)) = 1$ and $P_1(H) > P^H$, $\pi_2(H) = \pi_2(H, b(P_2(H)), P_2(H)) < \pi_2(H, 1, P_1(H))$, and so the high-quality firm would do better to select $P_1(H)$ in period two as well.

Consider next $P_t(L)$. Clearly, $P_t(L) = P^L$ or $P_t(L) = P_2(H)$ at any t . $P_1(L) = P^L$ and $P_2(L) = P_2(H)$ is impossible, since

$$\begin{aligned} \pi_2(L, b(P_2(H)), P_2(H)) \\ \geq \pi_2(L, b(P^L), P^L) \end{aligned}$$

implies $\pi_1(L, b(P_2(H)), P_2(H)) > \pi_1(L, 0, P^L)$, contradicting $P_1(L) = P^L$. Hence, $P_1(L) \geq P_2(L) \geq P^L$.

We now impose the additional assumption that X_2 is sufficiently large, taken here to mean $X_2 > \bar{X}$ and $P^\dagger(X_2) < \bar{P}(X_1)$. $X_2 > \bar{X}$ implies $P_2(L) = P^L$, while $P^\dagger(X_2) < \bar{P}(X_1)$ can be shown to imply $P_2(H) \in (\underline{P}(X_1), \bar{P}(X_1))$. The latter result gives $P_1(L) = P_2(H)$. Since $b(P_1(H)) = 1$ and $P_2(H) > P^L$, we thus have $P_1(H) > P_2(H) = P_1(L) > P_2(L) = P^L$.

REFERENCES

- Bagwell, Kyle, "Introductory Price as a Signal of Cost in a Model of Repeat Business," *Review of Economic Studies*, July 1987, 54, 365-84.
- , "Optimal Export Policy for a New-Product Monopoly," Northwestern University Discussion Paper 898, 1990.
- and Riordan, Michael H., "Equilibrium Price Dynamics for an Experience Good," Northwestern University Discussion Paper 705, 1986.
- and ———, "High and Declining Prices Signal Product Quality," Northwestern University Discussion Paper 808, 1989.
- Bernheim, B. Douglas and Whinston, Michael D., "Multimarket Contact and Collusive Behavior," *Rand Journal of Economics*, Spring 1990, 21, 1-26.
- Buzzell, Robert D., Nourse, Robert E. M., Matthews, John B. and Levitt, Theodore, *Marketing: A Contemporary Analysis*, New York: McGraw-Hill, 1972.
- Chan, Yuk-Shee and Leland, Hayne, "Prices and Qualities in Markets with Costly Information," *Review of Economic Studies*, October 1982, 49, 499-516.
- Cho, In-Koo, "A Refinement of Sequential Equilibrium," *Econometrica*, November 1987, 55, 1367-90.
- and Kreps, David M., "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, May 1987, 102, 179-221.
- Conlisk, John, Gerstner, Eitan and Sobel, Joel, "Cyclic Pricing by a Durable Goods Monopolist," *Quarterly Journal of Economics*, August 1984, 99, 489-505.
- Cooper, Russell and Ross, Thomas W., "Prices, Product Qualities and Asymmetric Infor-

- mation: The Competitive Case," *Review of Economic Studies*, April 1984, 51, 197-208.
- _____, and _____, "Monopoly Provision of Product Quality with Uninformed Buyers," *International Journal of Industrial Organization*, September 1985, 3, 439-49.
- Cr  mer, Jacques, "On the Economics of Repeat Buying," *Rand Journal of Economics*, Autumn 1984, 15, 396-403.
- Curry, David J. and Riesz, Peter C., "Prices and Price/Quality Relationships: A Longitudinal Analysis," *Journal of Marketing*, January 1988, 52, 36-51.
- Davis, Scott, *The Role of Price as a Signal of Product Quality in Monopolistic Markets*, Ph.D. Dissertation, Stanford University, 1989.
- Farrell, Joseph, *Prices as Signals of Quality*, Ph.D. Dissertation, Brasenose College, Oxford, 1980.
- _____, "Moral Hazard as an Entry Barrier," *Rand Journal of Economics*, Autumn 1986, 17, 440-9.
- Fertig, Doron, *Advertising as a Signal of Quality*, Ph.D. Dissertation, Northwestern University, 1988.
- Gabor, Andr   and Granger, Clive, "The Pricing of New Products," *Scientific Business*, August 1965, 3, 141-50.
- Gerstner, Eitan, "Do Higher Prices Signal Higher Quality?" *Journal of Marketing Research*, May 1985, 22, 209-15.
- Judd, Kenneth L., "The Law of Large Numbers with a Continuum of IID Random Variables," *Journal of Economic Theory*, February 1985, 35, 19-25.
- _____, and Riordan, Michael H., "Price and Quality in a New Product Monopoly," mimeograph, Stanford University, 1987.
- Kohlberg, Elon and Mertens, Jean-Francois, "On the Strategic Stability of Equilibria," *Econometrica*, September 1986, 54, 1003-38.
- Kreps, David M. and Wilson, Robert, "Sequential Equilibria," *Econometrica*, July 1982, 50, 863-94.
- Lazear, Edward P., "Retail Pricing and Clearance Sales," *American Economic Review*, March 1986, 76, 14-32.
- Milgrom, Paul and Roberts, John, "Price and Advertising Signals of Product Quality," *Journal of Political Economy*, August 1986, 94, 796-821.
- Monroe, Kent B., "Buyers' Subjective Perceptions of Price," in Harold H. Kassarian and Thomas S. Robertson, eds., *Perspectives in Consumer Behavior*, Glenview, IL: Scott, Foresman, 1973, 23-42.
- Ramey, Garey, "Moral Hazard, Signaling, and Product Quality," mimeograph, Stanford University, 1986.
- Riordan, Michael H., "Monopolistic Competition with Experience Goods," *Quarterly Journal of Economics*, May 1986, 101, 255-80.
- Shapiro, Carl, "Optimal Pricing of Experience Goods," *Bell Journal of Economics*, Autumn 1983, 14, 497-507.
- Stokey, Nancy L., "Rational Expectations and Durable Goods Pricing," *Bell Journal of Economics*, Spring 1981, 12, 112-28.
- Tellis, Gerard J. and Wernerfelt, Birger, "Competitive Price and Quality Under Asymmetric Information," *Marketing Science*, Summer 1987, 6, 240-53.
- Tirole Jean, *The Theory of Industrial Organization*, Cambridge, MA: MIT Press, 1988.
- Wolinsky, Asher, "Prices as Signals of Product Quality," *Review of Economic Studies*, October 1983, 50, 647-58.
- Zeithaml, Valerie A., "Consumer Perceptions of Price, Quality and Value: A Means-End Model and Synthesis of Evidence," *Journal of Marketing*, July 1988, 52, 2-22.

Striking for a Bargain Between Two Completely Informed Agents

By RAQUEL FERNANDEZ AND JACOB GLAZER*

This paper models the wage-contract negotiation procedure between a union and a firm as a sequential bargaining process in which the union must decide, in each period, whether or not to strike for the duration of that period. We show that there exist subgame-perfect equilibria in which the union engages in several periods of strikes prior to reaching a final agreement, although both parties are completely rational and fully informed. This has implications for other inefficient phenomena, such as tariff wars, debt negotiations, and wars in general. We characterize the set of equilibria, show that strikes can occur in real time, and discuss extensions of the model, such as lockouts and the possibility of multiple recontracting opportunities.

Economic theory has trouble explaining strikes.¹ As stated by Oliver Hart (1989 p. 25), "The difficulty is to understand why rational parties should resort to a wasteful mechanism as a way of distributing the gains from trade. Why could not both parties be made better off by moving to the final distribution of surplus immediately...and sharing the benefits from increased production?" A similar objection to developing a coherent theory of strikes is what John Kennan (1986 p. 1091) calls the "Hicks paradox," namely: "The main obstacle is that if one has a theory which predicts when a strike will occur and what the outcome will be, the parties can agree to this outcome in advance, and so avoid the costs of a strike. If they do this, the theory ceases to hold... If the parties are rational, it is difficult to see why they would fail to negotiate a Pareto optimal outcome."

This paradox has been resolved by resorting to informational imperfections, in par-

ticular, asymmetric information. Indeed, it is often thought that there are no other possible culprits for these inefficiencies.² David Card (1988 p. 1), for example, asserts that "It has long been recognized that any consistent theoretical model of strikes must appeal to some form of imperfect information." The basic idea underlying the asymmetric-information bargaining models developed in Anat Admati and Motty Perry (1987), Lawrence M. Ausubel and Raymond J. Deneckere (1989), Kalyan Chatterjee and Larry Samuelson (1987), Peter C. Cramton (1984), Drew Fudenberg et al. (1985), Sanford Grossman and Perry (1986), Hart (1989), Ariel Rubinstein (1985), and Joel Sobel and Ichiro Takahashi (1983) is that strikes, or delays in reaching agreement, are a signalling device. If a firm's profitability is unobservable by workers, then the willingness of a firm to delay agreement and therefore to forgo the output associated with such a delay serves as a signal of that firm's lower profits and allows a lower wage agreement to be reached. A high-profit firm would prefer to accept the higher wage agreement and obtain the revenue associated with pro-

*Fernandez: Boston University and NBER; Glazer: Boston University, Economics Department, 270 Bay State Rd., Boston, MA 02215. We thank Bob Rosenthal and two anonymous referees for helpful comments. The first author acknowledges financial support by NSF grant SES89-08390.

¹For a review of theories that attempt to explain strikes, see John Kennan (1986).

²Although, of course, bounded rationality could produce inefficient behavior. See Orley Ashenfelter and G. E. Johnson (1969) for a bargaining model in which only one side behaves optimally.

duction in those periods. Empirical work by Henry S. Farber and Max H. Bazerman (1989) and by Card (1988), however, casts some doubt on the ability of this kind of theory to explain reality. Moreover, in most asymmetric-information models, the Coase conjecture holds. That is, as the length of time separating bargaining periods becomes arbitrarily small, so does the real time of delay (see Faruk Gul and Hugo Sonnenschein [1988] for a rigorous discussion of this result).

Behind the assertion that imperfect information is the sole force driving strikes lies the implicit belief that, in the absence of informational asymmetries, bargaining between two parties is efficient. Both the cooperative- and the noncooperative-bargaining literature can be seen as lending support to that belief. The solution concepts of cooperative-bargaining theory, such as the Nash bargaining solution, assume Pareto-efficient outcomes. Moreover, the best-known examples of Rubinstein's (1982) noncooperative-bargaining model also produce unique and Pareto-efficient equilibria. In the case of fixed bargaining costs of c_i per period (where i indexes the name of the player, and with no discounting), if $c_1 = c_2$ it is possible to have inefficient equilibria emerge; for any $c_1 \neq c_2$, however, the subgame-perfect equilibrium is efficient.

Our paper's contribution is to show that strikes and other wasteful phenomena, such as wars, can result as equilibrium behavior within a framework of perfect rationality and complete information. Irrationality or informational asymmetries, while undoubtedly important factors in the explanation of many inefficient activities, are not necessary conditions for these to occur. Bargaining between two perfectly informed agents need not be efficient.

We develop a modified version of Rubinstein's (1982) bargaining model.³ As in

Rubinstein's model, the two agents—in our case, a union and a firm—are assumed to bargain sequentially over discrete time and a potentially infinite horizon. The union and firm alternate in making offers of wage contracts, which the other party is free to accept or reject. In our model, however, there is also an old wage contract (this is what is being renegotiated). This matters because, upon either party's rejection of a proposed wage contract, the union faces another decision: whether or not to strike in that period. If the union chooses to strike, it forgoes the wage that it would have received by not striking and instead working that period. That wage is assumed to be the one stipulated by the old contract. Thus, a decision to strike is costly to both parties. The union does not get paid, and the firm does not receive the revenue net of the wage bill. There is no uncertainty in this model, and agents possess complete information.

We show that there exist multiple subgame-perfect equilibria, some of which are Pareto-inefficient. The latter equilibria can take the following form: along the equilibrium play, the union makes very high wage offers, which the firm rejects. The firm, in turn, makes very low wage offers, which the union rejects. In every period in which an offer is rejected, the union strikes. This behavior continues for T periods, after which time an offer that lies somewhere between the high and low wage offers is both made and accepted. Despite the fact that reaching that same final agreement T periods earlier would be a Pareto improvement, we show that neither party will attempt to deviate from the equilibrium play behavior described above. Any attempt by one of the parties to deviate and reach an earlier agreement results in both the firm and the union thereafter playing an efficient equilibrium, but one which adversely affects the deviating party. Thus, we are able to answer the question posed in the first paragraph as to why it is that rational (and completely

³After completion of this paper, it was called to our attention that Hans Haller (1988) and Steinar Holden (1989) have developed models very similar to ours. Haller (1988 p. 16) however, concludes that there are no inefficient equilibria and therefore no strikes in

equilibrium, since "With complete information and rational players, bargaining is efficient." We show this conclusion to be incorrect.

informed) agents may engage in inefficient behavior.

The primary purpose of this paper is not so much to propose an alternative theory of strikes as to dispel a popular misconception concerning the necessity of asymmetric information for an explanation of this phenomenon. While our model has the less attractive feature that strikes occur only in some of the equilibria, it is also true that strikes in our model are not an artifact of the discrete-time bargaining framework: strikes can occur in real time; that is, they can be lengthy despite agents' ability to negotiate extremely rapidly. Furthermore, our model (or an extension of it) has as an implication that the range of parameter values that support strikes is greater in boom periods than in periods of recession. This is suggestive of the empirical finding that strikes tend to be procyclical.⁴ Other testable implications of our model include the specification of a range (given by a function of the firm's revenue in the case of no strike and by the union's wage in the status quo contract) in which strikes should not be observed.

The paper is organized as follows. In Section I, we set up the model. We discuss the efficient equilibria in Section II and the inefficient ones in Section III. In Section IV, we analyze some extensions of the model: we allow the firm to engage in lock-outs, and we examine the effect of multiple (predetermined) recontracting opportunities. Our conclusions are presented in Section V.

I. The Model

We consider the following situation: two parties—a union (of L identical workers hereafter normalized to equal 1) and its firm—have a contract that specifies the wage that a worker in the union is entitled to per day of work. This contract, however, has come up for renegotiation. The institutional mechanism governing contract renegotia-

tions is assumed to be as follows: the union and firm alternate in making wage offers over discrete time periods $t \in \{1, 2, \dots\}$. In each odd-numbered period (a period is taken for simplicity to be a day) the union proposes a wage contract x_t . The firm then responds (R_t) by either accepting the offer (Y) or rejecting it (N). If the firm accepts the offer, negotiations are over, and the newly agreed upon wage contract is assumed to hold thereafter (we later relax this assumption and allow contracts to be renegotiated any number of times). If the firm rejects the wage offer, the union must then make a decision S_t : to strike (s) or not to strike (ns). If the union decides not to strike that period, workers work and receive the old wage w_0 , $0 \leq w_0 \leq F$, specified by the preexisting contract, and the firm obtains the revenue F associated with the union's output minus the wage bill, that is, $F - w_0$. If the union decides to strike, workers forfeit their wage that period, and the firm does not earn $F - w_0$. Each party's payoff in this period is normalized to zero. After the union executes its decision S_t , time advances one period. In every even-numbered period, the firm offers the union a wage contract y_t . The union then responds (Q_t) by accepting (Y) or rejecting (N) the firm's proposal. Once again, acceptance implies that this new contract holds thereafter. Rejection of the offer, on the other hand, faces the union with the strike decision. The same rules as described previously govern the consequences of the strike decision. Once the decision is executed, time advances one period. Note that this bargaining process can potentially last an infinite amount of time. Figure 1 depicts the first two periods of the game.

The firm possesses a discount factor of $0 < \delta_f < 1$, and the union possesses a discount factor of $0 < \delta_u < 1$. The union's objective is to maximize workers' utility, the discounted sum of wage earnings,

$$\sum_{t=1}^{\infty} \delta_u^{t-1} w_t$$

and the firm's objective is to maximize the

⁴See Kennan (1986) for a discussion of the empirical work in this area.

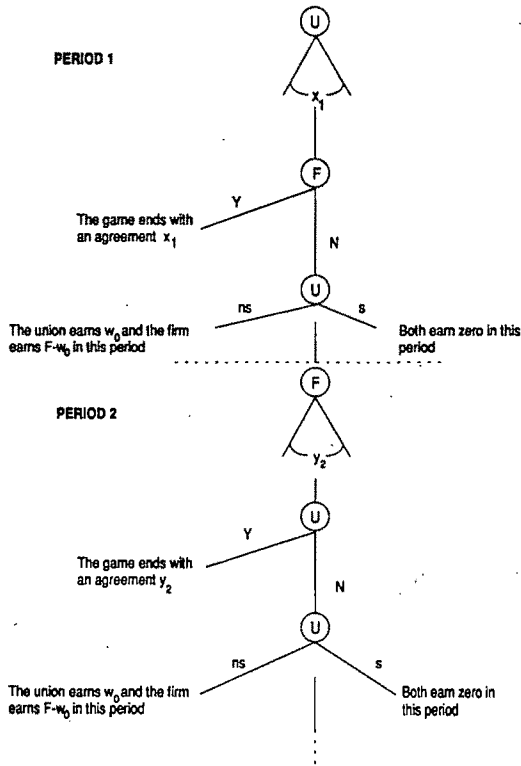


FIGURE 1. DIAGRAM OF THE FIRST TWO PERIODS OF THE GAME

discounted sum of profits,

$$\sum_{t=1}^{\infty} \delta_f^{t-1} (F - w_t).$$

Although the union is assumed to earn w_0 in the nonstrike periods prior to signing a new contract, it is also possible to view the negotiation process as including retroactive wage increases. This would not change any of our results, since what matters to the firm and to the union is the appropriately discounted value of earnings. Thus, a new wage contract w can be viewed as consisting partly of retroactive compensation and partly of wage increase.

We will be studying the subgame-perfect equilibria of the game described above. Subgame perfection is the natural refine-

ment of Nash equilibrium for a game with complete information, such as ours. Subgame perfection eliminates those equilibria based on "incredible" threats, that is, on threats that an agent would not be willing to carry out (they would be payoff-worsening for that player). That is, subgame-perfect-equilibrium strategies induce Nash equilibria in the game and in every proper subgame, including those subgames that will not be reached along the equilibrium play.

It is convenient to ask what the bargaining outcome would be if the union were committed to striking in every period in which it did not reach an agreement with the firm. As we will show, this is tantamount to assuming that the original wage contract does not exist, since w_0 is now no longer a possible cost of disagreement.

LEMMA 1: *If the union is committed to striking in every period in which there is a disagreement, then there is a unique subgame-perfect equilibrium to the bargaining game between the union and the firm. This equilibrium has agreement reached in the first period of negotiation and results in a wage contract of \bar{w} if bargaining commences in an odd-numbered period and has a contract of \bar{z} if bargaining commences in an even-numbered period, where*

$$\bar{w} = \frac{(1 - \delta_f) F}{1 - \delta_u \delta_f} \quad \bar{z} = \frac{\delta_u (1 - \delta_f) F}{1 - \delta_u \delta_f}.$$

PROOF:

See Rubinstein (1982) or Avner Shaked and John Sutton (1984).

Note that \bar{w} and \bar{z} are the solutions to Rubinstein's original bargaining game. The intuition underlying this result is that the union's commitment to strike in every period of disagreement transforms the game into Rubinstein's original bargaining model with both parties bargaining over a cake of size F . The fact that $\bar{w} > \bar{z}$ shows that the player who makes the first offer has an advantage in this kind of bargaining game.

II. Efficient Equilibria

In this section, we completely characterize the set of Pareto-efficient subgame-perfect equilibria. We first discuss three particular equilibria that are especially useful. One is the minimum wage contract that can be obtained in equilibrium, another is the maximum, and the third has the property that the union threatens to strike in each period in which an agreement is not reached.

LEMMA 2: *There is a subgame-perfect-equilibrium in which an agreement of w_0 is reached in the first period.*

PROOF:

The pair of subgame-perfect-equilibrium strategies given below generates a wage contract of w_0 in the first period. The union's strategy is never to strike (i.e., $S_t = \text{ns}$ for all t) and to offer $x_t = w_0$ in every odd-numbered t and, in every even-numbered t to reply to an offer y_t by

$$Q_t = \begin{cases} Y & \text{if } y_t \geq w_0 \\ N & \text{otherwise.} \end{cases}$$

The firm's strategy is to offer $y_t = w_0$ in every even-numbered t and, when t is odd, to reply to an offer x_t by

$$R_t = \begin{cases} Y & \text{if } x_t \leq w_0 \\ N & \text{otherwise.} \end{cases}$$

It is easy to check that these are subgame-perfect-equilibrium strategies.

Note that w_0 is the minimum wage contract that the union can receive, since it always has the option of working at the preexisting wage.⁵

⁵Furthermore, in a finite-horizon version of our model, w_0 is the sole subgame-perfect-equilibrium outcome; no strikes can occur. For a complete-information finite-horizon Rubinstein model (but with a commitment mechanism) capable of generating delays, see Chaim Fershtman and Daniel Seidmann (1990).

LEMMA 3: *If*

$$(1) \quad w_0 \leq \delta_u \bar{z}$$

there exists a subgame-perfect equilibrium in which an agreement of \bar{w} is reached in the first period.

PROOF:

A formal proof is contained in the proof of Theorem 2 below and is therefore omitted here.⁶

The following is an informal description of the strategies that generate the above equilibrium. The union offers the contract \bar{w} in every odd-numbered period, accepts any offer greater or equal to \bar{z} in every even-numbered period, and strikes in every odd-numbered period in which its request for \bar{w} is rejected and in every even-numbered period in which it is not offered at least \bar{z} . If, however, at some point, the union deviates from this rule, then the strategies call for both players to play thereafter according to the strategies described in Lemma 2. In other words, a deviation by the union is punished by having it accept the old wage contract of w_0 .

The maximum wage contract that the union can obtain, however, is not \bar{w} . This wage contract is established in the following lemma. Let

$$w' = \bar{w} + \delta_f w_0 (1 - \delta_u) (1 - \delta_u \delta_f)^{-1}$$

and

$$z' = \bar{z} + w_0 (1 - \delta_u) (1 - \delta_u \delta_f)^{-1}.$$

LEMMA 4: *If $w_0 \leq \delta_u z'$, there is a subgame-perfect equilibrium in which an agreement of w' is obtained in the first period. This is also the maximum wage contract that the*

⁶Theorem 2 provides a pair of strategies that produce \bar{w} as an outcome in some subgames. It is not difficult to see how they can be modified to generate \bar{w} as an agreement reached in the first period.

union can receive in any subgame-perfect equilibrium.

PROOF:

A proof that there exists a pair of subgame-perfect-equilibrium strategies that support w' as an equilibrium outcome in the first period is given in the proof of Theorem 1. The following is an informal description of the strategies that generate the above equilibrium. In odd-numbered periods, the union offers the contract w' and strikes if this offer is rejected. In even-numbered periods, the union accepts only offers that are greater than or equal to z' but never strikes. If, however, the union deviates from this rule at some point, then the strategies call for both players to play thereafter according to the strategies described in Lemma 2. In other words, a deviation by the union is punished by having it accept the old wage contract of w_0 .

We now show that w' is the maximum wage that the union can obtain in any subgame-perfect equilibrium. Suppose that it is not. Let $w^* > w'$ be the supremum over all wage agreements obtained in any subgame of any subgame-perfect equilibrium. Consider a subgame in which an agreement of $\hat{w} = w^* - \varepsilon$, $\varepsilon < \min[(w^* - w_0)(1 - \delta_u), (w^* - w')(1 - \delta_u\delta_f)]$, is reached. By hypothesis, at least one such subgame must exist. There are two cases to consider.

(i) Suppose this agreement occurs in a subgame in which the firm makes the offer. The following deviation is then profitable for the firm. Let the firm change its offer from \hat{w} to $\phi + \varepsilon'$, where $\phi = w_0(1 - \delta_u) + \delta_u w^*$ and $0 < \varepsilon' < (w^* - w_0)(1 - \delta_u) - \varepsilon$. Note that the union will accept an offer of $\phi + \varepsilon'$, since by rejecting this offer the most that the discounted value of its earnings can be is $w_0 + (1 - \delta_u)^{-1}\delta_u w^*$. However, the discounted value of $\phi + \varepsilon'$ is $w_0 + (1 - \delta_u)^{-1}(\delta_u w^* + \varepsilon')$. The firm gains since by construction, $\phi + \varepsilon' < \hat{w}$.

(ii) Suppose this agreement occurs in a subgame in which the union makes the offer. The following deviation is then profitable for the firm. Let the firm reject the union's offer and, in the following period, independently of whether the union chose

to strike in the previous period, offer the union a wage of $\phi + \varepsilon''$ where $0 < \varepsilon'' < [(1 - \delta_u\delta_f)(w^* - w') - \varepsilon]\delta_f^{-1}$. As shown in (i), the union will accept such an offer. The firm gains since, as can be shown by some algebraic manipulation and recalling that $w^* - \varepsilon > w'$, $F - \hat{w} < \delta_f(F - \phi - \varepsilon'')$.

Intuitively, the reason why this strategy of striking only in odd-numbered periods yields a greater wage contract than the strategy described in Lemma 3 of striking in every period, is that the first strategy creates an asymmetry in each party's costs of rejecting the other's offer. It is now more costly for the firm to reject the union's offer than it is for the union to reject the firm's offer, since rejection of the union's offer leads to a strike (with the consequent loss of profit for the firm), whereas the rejection of the firm's offer still allows the union to earn w_0 .

An alternative interpretation of w' is to note that w' can be written as

$$w' = w_0 + (1 - \delta_f)(F - w_0)(1 - \delta_u\delta_f)^{-1}.$$

That is, w' is equal to w_0 plus the solution to the original Rubinstein game in which the cake is of size $F - w_0$. By employing a strategy of striking only in odd-numbered periods, it is as though the players are bargaining over a cake of size $F - w_0$ and the union is already guaranteed a return of w_0 .

We now characterize the entire set of subgame-perfect-equilibrium wage contracts. Moreover, we show that all these contracts can be generated by Pareto-efficient subgame-perfect-equilibrium strategies.

THEOREM 1: *If $w_0 \leq \delta_u z'$, then any wage contract w such that $w_0 \leq w \leq w'$ can be generated as an equilibrium wage contract with agreement reached in the first period.*

PROOF:

We first introduce the following notation. Suppose that the game has reached period t . For every period $\tau < t$ let D'_τ be a function of the actions taken in that period such

that

$$D'_\tau = \begin{cases} d & \text{if } \tau \text{ is odd and } x_\tau > w'; \\ & \text{if } \tau \text{ is even and } y_\tau \geq z' \text{ but } Q_\tau = N; \text{ or} \\ & \text{if } S_\tau = ns \text{ and } \tau \text{ is odd} \\ nd & \text{otherwise.} \end{cases}$$

D'_τ indicates whether or not the union has deviated in period τ . (Note that, strictly speaking, D'_τ does not capture all possible deviations, since it ignores those offers by the union lower than w' .) If a deviation has occurred in period τ then $D'_\tau = d$; if not, then $D'_\tau = nd$. Similarly, suppose that the play has reached the last move of period t , at which point the union has to decide whether or not to strike. Let D'_t be a function of all actions taken in period t up to the strike decision such that

$$D'_t = \begin{cases} d & \text{if } t \text{ is odd and } x_t > w'; \text{ or} \\ & \text{if } t \text{ is even and } y_t \geq z' \text{ but } Q_t = N \\ nd & \text{otherwise.} \end{cases}$$

Let w be such that the $w_0 \leq w \leq w'$. Then, the following strategies constitute an equilibrium. The union's strategy is as follows:

$$x_1 = w$$

and for t odd and greater than one

$$x_t = \begin{cases} w_0 & \text{if } x_1 > w; \\ & \text{if } S_1 = ns; \text{ or} \\ & \text{if } D'_\tau = d \text{ for some } \tau, 1 < \tau < t \\ w' & \text{otherwise.} \end{cases}$$

When t is even the union's response is

$$Q_t = \begin{cases} Y & \text{if } y_t \geq z'; \\ & \text{if } y_t \geq w_0 \text{ and either } x_1 > w \text{ or } S_1 = ns; \\ & \text{or} \\ & \text{if } D'_\tau = d \text{ for some } \tau, 1 < \tau < t \\ N & \text{otherwise.} \end{cases}$$

Finally,

$$S_t = \begin{cases} ns & \text{if } x_1 > w; \\ & \text{if } S_1 = ns; \\ & \text{if } D'_\tau = d \text{ for some } \tau, 1 < \tau < t; \\ & \text{if } D_t = d; \text{ or} \\ & \text{if } t \text{ is even} \\ s & \text{otherwise.} \end{cases}$$

The firm's strategy is as follows: when t is even it offers

$$y_t = \begin{cases} w_0 & \text{if } x_1 > w; \\ & \text{if } S_1 = ns; \text{ or} \\ & \text{if } D'_\tau = d \text{ for some } \tau, 1 < \tau < t \\ z' & \text{otherwise.} \end{cases}$$

The firm's response in period 1 is

$$R_1 = \begin{cases} N & \text{if } x_1 > w \\ Y & \text{otherwise} \end{cases}$$

and in every odd period t , $t > 1$ it responds according to

$$R_t = \begin{cases} N & \text{if } x_t > w'; \\ & \text{if } x_t > w_0 \text{ and either } x_1 > w \text{ or} \\ & S_1 = ns; \text{ or} \\ & \text{if } D'_\tau = d \text{ for some } \tau, 1 < \tau < t \\ Y & \text{otherwise.} \end{cases}$$

Note that Theorem 1 implies that if w' is an equilibrium, then any w such that $w_0 \leq w \leq w'$ is also obtainable as an efficient subgame-perfect-equilibrium wage contract. Moreover, since w' is the maximum wage obtainable and w_0 is the minimum wage obtainable, this range describes the complete range of wage contracts that can be obtained as subgame-perfect equilibria.

III. Inefficient Equilibria

The purpose of this section is to show that, despite the existence of complete information, it is possible for bargaining to generate inefficient subgame-perfect equilibria. We limit our discussion to strikes that last for an uninterrupted T periods, although it is also possible to have periods of "peaceful" negotiations alternate with periods of strikes.

THEOREM 2: *If \hat{w} is such that*

$$(2) \quad (1 - \delta_f^{1-T})F + \delta_f^{1-T}\bar{z} \geq \hat{w} \geq \delta_u^{-T}w_0$$

then there is a subgame-perfect equilibrium in

*the play of which there is a strike of T periods followed by an agreement of \hat{w} .*⁷

PROOF:

See the Appendix for a formal presentation of the strategies.

We now provide an informal proof of the theorem, in which we limit ourselves to describing the strategies along the equilibrium path and to a discussion of the conditions sufficient for deviations not to occur. In each period prior to $T + 1$, the union makes a nonserious wage offer to the firm (i.e., the union offers a very high wage contract of F , which the firm rejects). In period $T + 1$, if this period is odd-numbered, the union offers $x_{T+1} = \hat{w}$; if it is even-numbered, the union accepts an offer $y_{T+1} = \hat{w}$. The union strikes in every period up to period $T + 1$. Prior to period $T + 1$, the firm also makes nonserious wage offers to the union (i.e., it offers the union very low wage contracts of w_0 , which the union rejects). In period $T + 1$, if this period is even-numbered, the firm offers $y_{T+1} = \hat{w}$; if it is odd-numbered, the firm accepts an offer $x_{T+1} = \hat{w}$.

It obviously would be a Pareto improvement if a settlement of \hat{w} were reached in any of the periods prior to $T + 1$. In fact, there exists a whole range of wage contracts that would be Pareto-improving if agreement on them were reached prior to $T + 1$. These potentially Pareto-improving deviations are blocked, however, by each party's response to deviations: attempts by the firm to "bribe" the union to reach a settlement earlier (by making wage offers such that the union prefers to accept the wage offered that period rather than wait until period $T + 1$ to obtain \hat{w}) are thwarted by having the strategies require both parties to play thereafter the equilibrium of Lemma 3 (which the union prefers to \hat{w}). That is, the union rejects the firm's offer, strikes, and in the following period offers the firm a wage contract of \bar{w} , which the firm then accepts. If, on the other hand, the union were to

deviate and attempt to reach an earlier settlement by offering the firm a wage contract that the latter preferred over obtaining an agreement of \hat{w} in period $T + 1$, or if the union simply decided not to strike, these deviations are thwarted by having the strategies require both parties to play thereafter the status quo wage equilibrium w_0 given in Lemma 2 (the best equilibrium outcome for the firm). That is, the firm would reject this offer, and next period it would offer the union a wage contract of w_0 , which the latter would then accept. The union's failure to strike likewise would lead to an agreement of w_0 in the following period. The actual \hat{w} agreement in period $T + 1$ is supported by the subgame-perfect-equilibrium strategies described in the proof of Theorem 1. That is, assuming no deviations have occurred prior to period $T + 1$, deviations at or after period $T + 1$ require both parties to play according to the strategies given in the proof of Theorem 1.

It is now easier to see how the conditions given in (2) are generated. The union always can obtain a wage contract of w_0 immediately, since the union always can choose not to strike and receive w_0 independently of any actions taken by the firm. Thus, in order for the union to be willing to strike for T periods, it should prefer to receive 0 for T periods, followed by a wage of \hat{w} thereafter, rather than to receive w_0 each period commencing in period 1. That is,

$$\delta_u^T \hat{w} \geq w_0$$

(i.e., \hat{w} must be sufficiently large).

If the firm were to attempt to reach an immediate settlement by offering the union $\bar{z} + \varepsilon$, $\varepsilon > 0$, it would not be subgame-perfect for the union to reject this offer, since next period it would receive \bar{w} , and $\bar{z} + \varepsilon > \delta_u \bar{w}$. Thus, in order for this deviation not to be performed by the firm, it must prefer to suffer the T periods of strike, followed by an agreement of \hat{w} , rather than to achieve an agreement of \bar{z} immediately (since \bar{z} is the lowest wage contract that can "bribe" the union). That is,

$$F - \bar{z} \leq \delta_f^{T-1} (F - \hat{w})$$

⁷It is actually possible to have a wider range for \hat{w} by substituting z' for \bar{z} in (2) and using the odd-period-only strikes equilibrium (described in Lemma 4).

or, rearranging terms,

$$(1 - \delta_f^{1-T})F + \delta_f^{1-T}\bar{z} \geq \hat{w}$$

(i.e., \hat{w} must be sufficiently small).

These are the only binding constraints on the size of \hat{w} , since any further deviations will only be less profitable than those just described.

Many models of bargaining under incomplete information have the feature that, even if an agreement is not reached immediately in equilibrium, the delay in reaching an agreement becomes arbitrarily small as the (exogenously given) time interval between two successive offers becomes arbitrarily small. In other words, when periods are short, allowing agents to alternate offers quickly, there is essentially no delay in reaching an agreement. This result is also known as the Coase conjecture.⁸

An important feature of our model is that shorter periods not only imply that agents alternate offers more rapidly, but also they accordingly shorten the length of a strike, given a decision to strike for a period. (If this were not the case, that is, if players could make offers and counteroffers quickly but the duration of a strike commitment remained unchanged, then all our previous results would go through trivially). However, it is easy to show that, even when the strike period becomes arbitrarily small, there exist equilibria with lengthy strikes. The reason for this is that strikes are not a signalling device for either player in our model. The union strikes because not doing so means that it will obtain a lower wage. The maximum length of strike time depends solely on the difference in payoffs that the union can obtain by striking as compared to not striking and not on any information revealed through delay; delay is no longer a mechanism by which different types separate themselves.

⁸Some notable exceptions are Admati and Perry (1987), Ausubel and Deneckere (1989), Peter C. Cramton (1989), and Hart (1989).

IV. Extensions

A. Lockouts

Thus far, we have examined equilibria that emerge when only the union is allowed to engage in actions other than offers, rejections, and acceptances. We would now like to ask how our results are affected if the firm is allowed to engage in lockouts. In order to simplify our computations, we assume that workers earn zero if they are locked out.⁹

If we start by examining a game in which only lockouts are feasible (the union is not allowed to strike), it is interesting to note that the equilibrium outcome obtained when the firm follows the strategy of locking out the union in every even-numbered period in which an agreement is not reached (i.e., a strategy analogous to the one described for the union in Lemma 4) now yields

$$\bar{w} = \frac{(1 - \delta_f)w_0}{1 - \delta_u\delta_f}$$

$$\bar{z} = \frac{\delta_u(1 - \delta_f)w_0}{1 - \delta_u\delta_f}$$

where \bar{w} is the wage contract obtained if the union commences the bargaining procedure and \bar{z} is obtained if the firm starts.¹⁰ This strategy again transforms the game into Rubinstein's (1982) original game, but now w_0 is the size of the cake being bargained over. Just as the odd-period-strike equilibrium allowed the union to bargain with the firm over the latter's right to earn $F - w_0$, the even-period-lockout strategy allows the firm to bargain with the union over the latter's right to work and receive w_0 . Thus,

⁹In reality, workers often receive unemployment compensation. This modification would not qualitatively affect our results.

¹⁰The condition for this to be a subgame-perfect equilibrium is

$$w_0 \geq F(1 - \delta_f)(1 - \delta_u\delta_f)[1 - \delta_u\delta_f - \delta_f(1 - \delta_f)]^{-1}.$$

this agreement yields the union a wage contract smaller than the status quo contract.

If we allow both lockouts and strikes (one can think of these decisions as following a rejection of an offer and occurring either simultaneously or sequentially), it is possible to have equilibria in which strikes, for example, alternate with lockouts along the equilibrium play before a final agreement is reached. These inefficient equilibria are sustained by having the strategies require the parties to play the w' equilibrium if the firm deviates (i.e., the best outcome for the union) and the \bar{w} equilibrium if the union deviates (the best outcome for the firm).

B. Multiple Contract Renegotiations

We now extend our model to allow for contracts that are repeatedly, potentially infinitely, renegotiated. We suppose that contracts are periodically renegotiated every M periods (periodicity is assumed for notational simplicity) after a contract has been established. All of the equilibrium outcomes described in the previous sections are also equilibrium outcomes in this modified setting [the strategies must now require that, after the first contract renegotiation (w^*) is concluded, the union will only accept and offer wage contracts of w^* or above and will never strike, and the firm will only accept and offer wage contracts of w^* and below]. In addition, we can now show that, unlike in the previous sections, the new equilibrium contract need not necessarily offer the union a wage greater than or equal to w_0 . As long as the union expects there to be a future contract in which its new wage will be sufficiently high so as to compensate it for the periods in which it worked for a lower wage, a wage contract lower than w_0 can be accepted as part of a subgame-perfect equilibrium. An example follows.

Example: Consider the following equilibrium play: in the first period, the union offers a wage contract of $p < w_0$, which the firm accepts. This contract comes up for renegotiation M periods later. Assuming $M + 1$ is odd, the union then offers a wage contract of $q > w_0$, which the firm accepts.

This outcome is supported by having the two parties play, as of the subgame following the deviation, the equilibrium strategies of Lemma 3 (that support \bar{w}) if the firm deviates and having them play the equilibrium strategies of Lemma 2 (that support w_0 if $t = 1$ and support p if $t > 1$) if the union deviates. In order for deviations not to be profitable, therefore, we must have the union prefer the equilibrium outcome to obtaining a wage contract of w_0 forever, that is,

$$(q - p)\delta_u^M + p \geq w_0.$$

Also, if the firm rejects the union's offer of p , the union must prefer to strike and obtain a wage contract of \bar{z} next period to working that period and thereafter for a wage of w_0 . Hence,

$$\delta_u \bar{z} \geq w_0.$$

V. Conclusion

This paper has shown that bargaining between two agents may be inefficient even if both parties are completely rational and fully informed. In our specific case of a union and a firm negotiating a new wage contract, we have shown that this process may involve periods of strikes. Hence, neither bounded rationality nor incomplete information is a necessary condition for a consistent theory of strikes. The length of time for which the union can strike depends on the status quo wage—the wage specified by the preexisting contract—and on the profitability of the firm. The lower the status quo wage and the more profitable the firm, the greater the maximum length of time for which the union may strike in equilibrium. The ability of the union to strike, even in those equilibria in which the union does not actually strike along the equilibrium play, can only improve the union's position at the bargaining table. Thus, the union's threat to strike may be credible despite the cost to the union of carrying out such a threat. Furthermore, even if the time separating bargaining periods becomes arbitrarily small, strikes can still occur in real time (i.e., lengthy strikes are still possible).

We have shown that our model can be extended to include multiple recontracting opportunities and the ability of the firm to engage in lockouts. Another interesting extension would be to include uncertainty in the form of shocks to the firm's revenue function through technology or price changes. If these shocks were perfectly observable to all parties, contracts could be renegotiated in the event of a shock. The range of parameter values that permit inefficient equilibria would be greater for positive shocks than for negative ones, which is suggestive of the empirical finding that strikes are procyclical.

Finally, our paper's main result—that bargaining between two parties may result in inefficient outcomes—may justify the existence of Pareto-inferior phenomena other than strikes. Observed inefficiencies in the international arena (e.g., the existence of tariff wars or protracted debt negotiations interspersed with periods of debt moratoria) may also be explained by our model. In particular, our model can offer an explanation for why two completely rational countries may engage in war although their disagreements could be settled via the much less costly process of diplomacy.

APPENDIX

We provide a pair of subgame-perfect equilibrium strategies that generate strikes for T periods followed by an agreement of \hat{w} in period $T+1$. We assume here that T is even-numbered.

For every t , $1 \leq t \leq T+1$, let AD_t be a function of the history of play up to (but not including) period t such that

$$AD_1 = nd$$

and for $t > 1$,

$$AD_t = \begin{cases} nd & \text{if for every } \tau, 1 \leq \tau < t, S_\tau = s; \text{ for every odd } \tau, x_\tau = F; \text{ and, for every even } \tau, y_\tau = w_0 \\ df & \text{if there exists some even } \tau' < t \text{ such that } AD_{\tau'} = nd \text{ but } y_{\tau'} > w_0, S_{\tau'} = s, \text{ and } D_\tau = nd \text{ for all } \tau, \tau' < \tau < t \\ du & \text{otherwise} \end{cases}$$

where D_τ is the equivalent of D'_τ with the substitution of \bar{w} for w' , \bar{z} for z' , and $S_\tau = ns$ for all $\tau < t$.

The function AD_t indicates whether, prior to period t , any of the players deviated from equilibrium play and identifies this player. If $AD_t = nd$, no deviation has occurred; if $AD_t = df$, the firm has deviated and the union has not; and if $AD_t = du$, the union has deviated.

For every t , $1 \leq t < T+1$, let DD_t be a function from the history of play at period t such that

$$DD_t = \begin{cases} d & \text{if } AD_t = nd \text{ and } t \text{ is odd and } x_t < F; \\ & \text{if } AD_t = df \text{ and either } t \text{ is odd and } x_t > \bar{w} \text{ or } t \text{ is even and } y_t \geq \bar{z} \text{ but } Q_t = N; \text{ or} \\ & \text{if } AD_t = du \\ nd & \text{otherwise.} \end{cases}$$

DD_t indicates whether or not the union has deviated in or prior to period t before its decision of whether or not to strike.

For $t > T+1$, let BD_t be a function of the history of play up to (but not including) period t , such that

$$BD_t = \begin{cases} df & \text{if } AD_{T+1} = df \text{ and } D_\tau = nd \text{ for all } \tau, T+1 \leq \tau < t; \text{ or} \\ & \text{if } AD_{T+1} = nd, x_{T+1} \leq \hat{w}, S_{T+1} = s, \text{ and } D_\tau = nd \text{ for all } \tau, T+1 < \tau < t \\ du & \text{otherwise.} \end{cases}$$

The function BD_t indicates whether the union or only the firm has deviated from the equilibrium rule.

The union's strategy is

$$x_1 = F$$

and in every odd-numbered t , $1 < t < T+1$, it offers

$$x_t = \begin{cases} F & \text{if } AD_t = nd \\ \bar{w} & \text{if } AD_t = df \\ w_0 & \text{otherwise.} \end{cases}$$

In period $T + 1$, it offers

$$x_{T+1} = \begin{cases} \hat{w} & \text{if } AD_{T+1} = \text{nd} \\ \bar{w} & \text{if } AD_{T+1} = \text{df} \\ w_0 & \text{otherwise.} \end{cases}$$

and for every odd-numbered t , $t > T + 1$,

$$x_t = \begin{cases} \bar{w} & \text{if } BD_t = \text{df} \\ w_0 & \text{otherwise.} \end{cases}$$

For $t < T + 1$, the union's response is

$$Q_t = \begin{cases} Y & \text{if } y_t \geq \bar{z}; \text{ or} \\ & \text{if } y_t \geq w_0 \text{ and } AD_t = \text{du} \\ N & \text{otherwise.} \end{cases}$$

For $t > T + 1$, the union's response is

$$Q_t = \begin{cases} Y & \text{if } y_t \geq \bar{z}; \text{ or} \\ & \text{if } y_t \geq w_0 \text{ and } BD_t = \text{du} \\ N & \text{otherwise.} \end{cases}$$

For $t < T + 1$, the union's strike decision is

$$S_t = \begin{cases} \text{ns} & \text{if } DD_t = d \\ s & \text{otherwise} \end{cases}$$

and in period $T + 1$,

$$S_{T+1} = \begin{cases} \text{ns} & \text{if } AD_{T+1} = \text{du}; \\ & \text{if } AD_{T+1} = \text{df} \text{ but } x_{T+1} > \bar{w}; \text{ or} \\ & \text{if } AD_{T+1} = \text{nd} \text{ but } x_{T+1} > \hat{w} \\ s & \text{otherwise.} \end{cases}$$

For every $t > T + 1$,

$$S_t = \begin{cases} \text{ns} & \text{if } BD_t = \text{du}; \text{ or} \\ & \text{if } BD_t = \text{df} \text{ but } D_t = d \\ s & \text{otherwise.} \end{cases}$$

where D_t is the equivalent of D'_t with the substitution of \bar{w} for w' and \bar{z} for z' .

The firm's strategy is as follows: when t is even and $t < T + 1$, it offers

$$y_t = \begin{cases} \bar{z} & \text{if } AD_t = \text{df} \\ w_0 & \text{otherwise} \end{cases}$$

and when $t > T + 1$, it offers

$$y_t = \begin{cases} \bar{z} & \text{if } BD_t = \text{df} \\ w_0 & \text{otherwise.} \end{cases}$$

When t is odd and $t < T + 1$, the firm's response is

$$R_t = \begin{cases} Y & \text{if } x_t \leq w_0; \text{ or} \\ & \text{if } x_t \leq \bar{w} \text{ and } AD_t = \text{df} \\ N & \text{otherwise} \end{cases}$$

and in period $T + 1$, it responds

$$R_{T+1} = \begin{cases} N & \text{if } x_{T+1} > \bar{w}; \\ & \text{if } x_{T+1} > \hat{w} \text{ and } AD_{T+1} = \text{nd}; \text{ or} \\ & \text{if } x_{T+1} > w_0 \text{ and } AD_{T+1} = \text{du} \\ Y & \text{otherwise.} \end{cases}$$

In every odd-numbered t , $t > T + 1$, the firm responds according to

$$R_t = \begin{cases} N & \text{if } x_t > \bar{w}; \text{ or} \\ & \text{if } x_t > w_0 \text{ and } BD_t = \text{du} \\ Y & \text{otherwise.} \end{cases}$$

REFERENCES

- Admati, Anat R. and Perry, Motty, "Strategic Delay in Bargaining," *Review of Economic Studies*, July 1987, 54, 343-64.
- Ashenfelter, Orley and Johnson, G. E., "Bargaining Theory, Trade Unions and Industrial Strike Activity," *American Economic Review*, March 1969, 59, 35-49.
- Ausubel, Lawrence M. and Deneckere, Raymond J., "Reputation in Bargaining and Durable Good Monopoly," *Econometrica*, May 1989, 57, 511-32.
- Card, David, "Strikes and Wages: A Test of a Signalling Model," NBER Working Paper No. 2550, April 1988.
- Chatterjee, Kalyan and Samuelson, Larry, "Bargaining with Two-Sided Incomplete Information: An Infinite Horizon Model with Alternating Offers," *Review of Economic Studies*, April 1987, 54, 175-92.

- Cramton, Peter C., "Bargaining with Incomplete Information: An Infinite Horizon Model with Continuous Uncertainty," *Review of Economic Studies*, October 1984, 51, 579-93.
- _____, "Strategic Delay in Bargaining with Two-Sided Uncertainty," working paper, Yale School of Management, 1989.
- Farber, Henry S. and Bazerman, Max H., "Divergent Expectations as a Cause of Disagreement in Bargaining: Evidence from a Comparison of Arbitration Schemes," *Quarterly Journal of Economics*, November 1989, 104, 99-120.
- Fershtman, Chaim and Seidmann, Daniel J., "Deadline Effects and Inefficient Delay in Bargaining With Endogenous Commitment," working paper, Tel Aviv University, 1990.
- Fudenberg, Drew, Levine, David and Tirole, Jean, "Infinite Horizon Models of Bargaining with One-Sided Incomplete Information," in Alvin Roth, ed., *Bargaining with Incomplete Information*, Cambridge, U.K.: Cambridge University Press, 1985, 73-98.
- Grossman, Sanford J. and Perry, Motty, "Sequential Bargaining Under Asymmetric Information," *Journal of Economic Theory*, June 1986, 39, 97-119.
- Gul, Faruk and Sonnenschein, Hugo, "On Delay in Bargaining with One-Sided Uncertainty," *Econometrica*, May 1988, 56, 601-11.
- Haller, Hans, "Wage Bargaining as a Strategic Game," mimeo, Virginia Polytechnic Institute, October 1988.
- Hart, Oliver, "Bargaining and Strikes," *Quarterly Journal of Economics*, February 1989, 104, 25-44.
- Holden, Steinar, "Non-Cooperative Wage Bargaining," mimeo, London School of Economics, May 1989.
- Kennan, John, "The Economics of Strikes," in O. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*, New York: Elsevier, 1986, 1091-137.
- Rubinstein, Ariel, "Perfect Equilibrium in a Bargaining Model," *Econometrica*, January 1982, 50, 97-109.
- _____, "A Bargaining Model with Incomplete Information About Time Preferences," *Econometrica*, September 1985, 53, 1151-72.
- Shaked, Avner and Sutton, John, "Involuntary Unemployment as a Perfect Equilibrium in a Bargaining Model," *Econometrica*, September 1984, 52, 1351-64.
- Sobel, Joel and Takahashi, Ichiro, "A Multi-Stage Model of Bargaining," *Review of Economic Studies*, July 1983, 50, 411-26.

An Experimental Analysis of Strikes in Bargaining Games with One-Sided Private Information

By ROBERT FORSYTHE, JOHN KENNAN, AND BARRY SOPHER*

We study two-player pie-splitting games in which one player knows the pie and the other knows only its probability distribution. We compare treatments in which incentive-efficient strikes (disagreements) are possible with alternatives in which efficiency forbids strikes. We find that incentive-efficiency is very helpful in explaining when strikes occur. There is also evidence of substantial heterogeneity in the subjects' altruism and in their risk preferences. This means that the common-knowledge assumptions of game theory cannot be controlled in experiments; but in our experiments the main theoretical conclusions seem robust to violations of these assumptions. (JEL 026, 215)

Recent theoretical analyses of strikes have emphasized the potential importance of private information in rationalizing the apparent inefficiencies associated with strike activity. The basic idea is that although strikes are not Pareto optimal *ex post* they may be Pareto optimal *ex ante*, in the sense that every alternative leaves one of the bargainers worse off in some contingency which cannot be ruled out on the basis of information that is common knowledge.

There are two serious difficulties with private-information bargaining models. First, there is an obvious problem in extracting empirical implications, since some of the relevant data, by definition, are not observable. Second, even simple models present formidable technical problems, so that it is necessary to make assumptions that are not literally credible in order to obtain results.

For example, strategic analyses proceed on the assumption that the timing of offers and counteroffers is governed by a procedure that is imposed arbitrarily.

A sensible response to these difficulties is to attempt to build simple models that give reasonably accurate descriptions of behavior observed in experiments, where the effects of private information can be measured directly. A particular goal of this work is to identify predictions that are robust with respect to the simplifying assumptions used in theoretical modeling.

This paper describes the result of a bargaining experiment involving a set of pie-splitting games between two players. The pie is a sum of money which takes on one of two values (good and bad) with known probabilities. One player knows the realized value of the pie, while the other knows only the probabilities of the two pie sizes. The players do not meet face-to-face, but they can negotiate by passing nonbinding handwritten messages to each other and by making binding offers to split the pie in a particular way. If agreement is not reached within 10 minutes, then neither bargainer gets anything.

If the uninformed bargainer has the right to issue an ultimatum in this game, as in the analysis of Drew Fudenberg and Jean Tirole (1983), then he can obtain the bad pie for sure, or he can obtain the good pie when it is there and zero otherwise. A self-

*Forsythe and Kennan: Department of Economics, University of Iowa, Iowa City, IA 52242; Sopher: Department of Economics, Rutgers University, New Brunswick, NJ 08903. We thank Martin Browning, Russell Cooper, Roger Myerson, and the participants in seminars at Princeton University, the University of Chicago, Carleton University, McMaster University, and the Hoover Institution for helpful comments. Portions of the paper were completed while Kennan was a National Fellow at the Hoover Institution. Martin Sefton provided valuable research assistance. This research was funded in part by the National Science Foundation, under grants SES-8510539, SES-8607771, and SES-8910930.

ish and risk-neutral uninformed bargainer will demand the good pie if his expected earnings from doing so exceed the size of the bad pie. We call this the *strike condition*: if it holds, the bargainers will fail to agree whenever the bad pie is drawn. In our experiment, however, no one has the right to make all-or-nothing demands. An alternative bargaining structure, proposed by Myerson (1984), allows the bargainers to negotiate over the set of direct mechanisms, which specify whether a strike should occur and how much each bargainer should get, as a function of the informed bargainer's announcement of the pie size. In this context, there is a large class of (interim incentive-) efficient mechanisms that do not involve a strike: any mechanism that gives the uninformed bargainer a fixed amount will do, provided each player gets a nonnegative payoff. If the strike condition holds, then there are also efficient mechanisms involving a strike in the bad state. Further, the incidence of strikes can be predicted by adopting a *random-dictator* (RD) axiom, which asserts that the bargainers will agree to a fair mixture of the mechanisms that each would unilaterally impose.

Both of these theoretical models suggest that strikes should occur more often in our experiment when the strike condition holds and when the pie is small. This prediction worked well, even though our experimental setting did not correspond exactly to either model. We also found, however, that strikes sometimes occurred in the good state, and also in the bad state, when the strike condition did not hold. Moreover, the two games in the experiment in which the strike condition was on gave quite different strike-incidence results. The RD axiom implies that strike incidence should depend only on the strike condition and not on any other parameters: this prediction failed in our experiment.

We also conducted an additional experiment designed to test the RD axiom directly and to calibrate the results of the bargaining experiment. In the RD experiment, both the informed and uninformed bargainers submitted all-or-nothing demands specifying the amount that the uninformed player

would receive. One of these demands was randomly chosen, and both players were paid accordingly if the selected demand was feasible (i.e., did not exceed the size of the pie). In these games, the informed bargainer should always keep the whole pie, while the uninformed bargainer should demand the good pie if the expected earnings from doing so exceed the size of the bad pie. Thus, there should always be a strike whenever the strike condition is on, the pie is small, and the uninformed bargainer's demand is chosen.

Comparing the RD games with the bargaining games, we found that the strike-incidence results were remarkably close. The RD results suggest that the unexplained variation in strike incidence in the bargaining games is due, to some extent, to violations of the risk-neutrality assumption. There is strong evidence of heterogeneity in the subject population, with a majority of risk-neutral types, but with substantial minorities of both risk-averse and risk-seeking types. Further, while the typical player was selfish, there was some altruistic behavior, which had a substantial effect on the outcomes.

In a series of experiments initiated by Alvin Roth, Michael Malouf, and Keith Murnighan (1979),¹ it has been found that the incidence of strikes in bargaining games tends to be related to the distribution of private information that is theoretically irrelevant. In these experiments, players bargained over lottery tickets that determined the probability of winning a prize, which was different for each player. If each player maximizes expected utility, knowledge of the prizes should be irrelevant in the bargaining game, and since both players know everything that is relevant to the game, there is no reason for strikes to occur. In fact, however, it was found that some strikes invariably occurred, and strike incidence was higher when it was not commonly known

¹See also Roth and Malouf (1979), Malouf and Roth (1981), Roth and Murnighan (1982), and Roth and Francoise Schoumaker (1983); for a review of this work, see Roth (1988).

that the player with the smaller prize knew both prizes. The main difference between this line of research and the experiments reported here is that we focus on the effect of private information that is relevant, according to the theory, because of incentive-compatibility constraints. In our setting, the theory includes strikes, and we can test whether the observed variations in strike incidence correspond to the theoretical predictions.

I. Efficient Bargaining Mechanisms

It is not possible to provide a complete strategic analysis of the bargaining games we conducted, since players could freely exchange messages. Instead, we consider two models that are abstractions of these games. The first is an ultimatum game, that is, a game in which one player can make an all-or-nothing demand which the other player must accept or reject. The second is a model in which players bargain over mechanisms. Both models lead to the same condition for efficient strikes.

We consider a class of pie-splitting games between two players, in which one player has private information. The pie is a sum of money, π , which takes the values π_g (good) with probability p and π_b (bad) with probability $1 - p$. One player already knows the realization of π when negotiations begin, while the other knows only the probability distribution from which π has been drawn. A fixed period of time is allowed for negotiation, and if agreement is not reached during this time, then the pie is lost. The period is short enough so that delaying an agreement is not costly (provided that agreement is reached before the deadline). It is also assumed that both players are selfish expected-income maximizers.

This is a simple example of a class of games that have been studied extensively in the bargaining literature.² A basic result is

that if the uninformed player has the ability to make nonnegotiable demands, then he will demand π_g (play "tough" in the language of Fudenberg and Tirole [1983]) if the condition $p\pi_g > \pi_b$ holds, and otherwise he will demand π_b (play "soft"). This is the strike condition: if it holds, the bargainers will fail to agree whenever the bad pie is drawn. This result is sensitive to the rules of bargaining, however: if the informed player could make nonnegotiable demands, for example, he would keep the pie in either state, and strikes would not occur. More generally, if neither player can commit to a particular bargaining position, then the outcome depends on each player's cost of delay, on the rules governing the timing of offers and counteroffers, and on which of many possible equilibria is reached. Still, a reasonable conjecture is that if the strike condition does not hold then strikes will not occur in equilibrium.

Because of the unconstrained messages which pass between the players, our bargaining game is too complicated to allow a detailed strategic analysis. However, by the revelation principle, any Nash equilibrium of this game is equivalent to some direct mechanism, which specifies whether a strike should occur and how much each bargainer should get as a function of the informed bargainer's announcement of the size of the pie. We will argue that if the equilibrium is not incentive-efficient then the uninformed bargainer will make an offer that the informed bargainer cannot afford to refuse. Thus, strikes will not occur unless they can be construed as part of an incentive-efficient mechanism.

Following Roger Myerson (1979, 1984), we analyze efficiency by assuming that the bargainers negotiate inscrutably over the set of direct mechanisms.³ We consider mechanisms that work as follows. The informed bargainer first announces that the state is good or bad. The uninformed bargainer can

²See, for example, Joel Sobel and Ichiro Takahashi (1983), Fudenberg and Tirole (1983), and Peter Cramton (1984). A critical review of this literature, as it pertains to the interpretation of strike data, appears in Kennan and Robert Wilson (1989).

³A detailed discussion of this model was presented in Kennan (1986), and extensions to the case of two-sided private information were discussed in Kennan (1987).

then reduce the pie by a known fraction $1 - \gamma$ (interpreted as the probability of disagreement or as the duration of a strike), leaving $\gamma_g \pi$ or $\gamma_b \pi$, according to the announcement made by the informed bargainer. The uninformed bargainer receives x_g or x_b , and the informed bargainer receives the rest of the pie. Thus, a mechanism μ is defined by four parameters (γ_j and x_j for $j = b, g$). A crucial assumption here is that the players can reach an agreement that commits them to implement a mechanism, even though one or both of them might wish to tear up the agreement immediately after it has been signed.

Incentive compatibility requires that it is optimal for the informed bargainer to tell the truth. The payoffs and the resulting truth-telling conditions can be tabulated as follows:

Valuation	Truth \geq Lie
Good	$\gamma_g \pi_g - x_g \geq \gamma_b \pi_g - x_b$
Bad	$\gamma_b \pi_b - x_b \geq \gamma_g \pi_b - x_g$

Thus, a mechanism is incentive-compatible (IC) if

$$(1) \quad (\gamma_g - \gamma_b) \pi_b \leq x_g - x_b \leq (\gamma_g - \gamma_b) \pi_g.$$

An IC mechanism is feasible if it is individually rational (IR), in that it gives nonnegative expected payoffs to each player in each state.

Incentive-efficiency⁴ is defined in terms of a Pareto ordering (over the set of feasible mechanisms) that respects the preferences of all types of all players, even though only one type of each player is actually present. In our context, this means that a mechanism is efficient if it is Pareto optimal for the set of three agents including the uninformed player, the informed player with the good pie, and the informed player with the bad pie.

We concentrate on the implications of the efficiency criterion for the incidence of "strikes," defined as a failure of the bar-

gainers to agree on a division of the pie. It is easy to show that strikes in the good state are not efficient: $\gamma_g = 1$. Indeed, $\gamma_g \geq \gamma_b$ by the IC constraints, so if $\gamma_g = 1 - \delta < 1$, we can define a new mechanism μ^* with $\gamma_g^* = 1$, $\gamma_b^* = \gamma_b + \delta$, $x_g^* = x_g$, and $x_b^* = x_b$. The mechanism μ^* does not affect the uninformed bargainer's expected payoff, but it increases the informed bargainer's payoff by $\delta \pi_g$ and $\delta \pi_b$ in the good and bad states, respectively, so the original mechanism cannot be efficient.

If γ_b can be increased without violating the incentive-compatibility constraint, the uninformed bargainer is unaffected, as is the informed bargainer in the good state, while the informed bargainer is made better off in the bad state. Thus, efficiency requires that the right-hand inequality in the incentive-compatibility constraint must hold with equality, that is,

$$(2) \quad x_g - x_b = (1 - \gamma_b) \pi_g.$$

We now show that the condition $p \pi_g > \pi_b$, which indicates that the uninformed bargainer would play tough in the noncooperative game, also determines whether efficient strike mechanisms exist in the cooperative game. Consider a mechanism $\mu = (\gamma_g, x_g, \gamma_b, x_b)$ which satisfies (2), with $\gamma_b < 1$. If strikes are not efficient, there is another mechanism which strictly dominates μ . Call this mechanism $\mu^* = (1, x_g^*, \gamma_b + \delta, x_b + d)$, where x_g^* is constructed to satisfy (2). Let ΔV_g , ΔV_b , and ΔU denote the resulting changes in expected payoffs to the informed bargainer in the good and bad states and to the uninformed bargainer. Then,

$$\Delta V_g = \pi_g \delta - d$$

$$\Delta V_b = \pi_b \delta - d$$

$$\Delta U = -p \pi_g \delta + d.$$

If all three of these changes are nonnegative and one is positive, then δ is not zero, and

$$\pi_g \delta \geq d \geq p \pi_g \delta$$

$$\pi_b \delta \geq d \geq p \pi_g \delta.$$

⁴This is "interim incentive efficiency," in the language of Bengt Holmstrom and Myerson (1983).

The first set of inequalities implies that δ is nonnegative, and (since δ is not zero) the second set then implies $\pi_b > p\pi_g$. In this case, strikes cannot be efficient, since values of δ and d can be chosen such that μ^* dominates μ . If the strike condition $p\pi_g > \pi_b$ holds, however, efficient strike mechanisms exist; in fact, any individually rational mechanism that satisfies (2) is efficient in this case. This also means that even when the strike condition holds there is a large class of efficient mechanisms that do not involve a strike, namely the class of IR mechanisms which give the uninformed bargainer a transfer that does not depend on the state.

Strike Incidence

So far, we have identified a necessary condition for efficient strikes, but we have not predicted the actual incidence of strikes when this condition holds. In fact, given that efficient strike mechanisms exist, the incidence of strikes depends on which mechanism is selected. Again following Myerson (1984), a sharp prediction can be made by imposing a random-dictator (RD) axiom, which asserts that the bargainers will agree to a fair mixture of the mechanisms that each would impose if given dictatorial power over the set of feasible mechanisms.

The RD axiom implies that strikes will occur with probability $\frac{1}{2}$ when the strike condition is satisfied and the bad pie is drawn. Thus, when $p\pi_g > \pi_b$ the uninformed bargainer would impose the mechanism $\mu = (1, \pi_g, 0, 0)$, which is an all-or-nothing demand for the big pie. When $p\pi_g < \pi_b$ the uninformed bargainer prefers π_b with certainty, so he will impose the mechanism $\mu = (1, \pi_b, 1, \pi_b)$, which is a fixed-transfer mechanism with no strikes. If the informed bargainer is the dictator, he will impose the mechanism $\mu = (1, 0, 1, 0)$, in which the uninformed bargainer always receives 0 and strikes never occur.

Randomizing equally between these two extremes gives the mechanisms

$$\mu = (1, \frac{1}{2}\pi_g, \frac{1}{2}, 0) \text{ when } p\pi_g > \pi_b$$

and

$$\mu = (1, \frac{1}{2}\pi_b, 1, \frac{1}{2}\pi_b) \text{ when } p\pi_g < \pi_b.$$

Thus, when the strike condition is on, this solution predicts that the incidence of strikes will be 50 percent when the pie is small, for all values of p , π_g , and π_b that satisfy $p\pi_g > \pi_b$. When the pie is big or when the strike condition is off, no strikes should occur.⁵

To sum up, the analysis of efficient bargaining mechanisms leads to clear predictions about the incidence of strikes. If a mechanism is not efficient, then either player can (without leaking information) propose an alternative that improves the outcome for both players. This implies that, if efficient strike mechanisms do not exist, then strikes will not occur. Thus, the theory predicts that strikes will occur only if the condition $p\pi_g > \pi_b$ holds, and then only if the bad pie is drawn. Further, imposing the RD axiom leads to the prediction that strikes should occur with probability $\frac{1}{2}$ when the strike condition is on and the pie is small.

II. The Experiments

A. Choice of Parameters

We used four different parameter sets, as shown in Table 1. Game I was run twice in the bargaining experiment; in game Ia there were two sessions, and in game Ib there were three sessions, as described below. In games I and II the strike condition is on, and in games III and IV the condition is off. Some early results led us to conjecture that strikes might be averted if the informed player could afford to offer half of the big pie, even when the small pie had actually been drawn. Thus, we chose parameters such that in games I and III the small pie is less than half of the big pie, while the opposite is true in games II and IV. In all of the games except Ib the participants played 10 separate games; in Ib the participants played 15 separate games.

⁵This solution is the same as Myerson's (1984) neutral bargaining solution for the simple bargaining games we consider here.

TABLE 1—PARAMETERS

Parameter set	p	π_g	π_b	$p\pi_g > \pi_b?$	Number of games played
Ia	$\frac{1}{2}$	6.00	1.00	yes	10
Ib	$\frac{1}{2}$	6.00	1.00	yes	15
II	$\frac{3}{4}$	3.90	2.30	yes	10
III	$\frac{1}{2}$	4.20	2.80	no	10
IV	$\frac{1}{4}$	6.80	2.40	no	10

B. Procedures

The experimental subjects were students recruited from undergraduate classes in economics at the University of Iowa. Each subject played 10 or 15 separate games against different anonymous opponents, with the same parameters (p, π_g, π_b) in each game. The experiments were conducted in sessions of five games each, held on two or three consecutive evenings. To encourage attendance at the subsequent sessions, earnings were withheld until the final session was completed. To eliminate incentives for reputation formation, participants were told nothing about their opponents' behavior in previous games. The participants were given no opportunity to see their opponents, before, during, or after the experiment.

We used two connecting rooms, each of which could accommodate up to seven subjects. Each participant alternated between being informed and uninformed in successive periods: the individuals in room A were uninformed in odd-numbered periods and informed in even-numbered periods, with the opposite designations in room B. The work of K. Binmore, A. Shaked, and J. Sutton (1984, 1985) suggests that allowing experimental subjects to play both sides of a bargaining game tends to eliminate an apparent inclination toward "fair" outcomes. On the other hand, each subject faced just one parameter set and, thus, had ample opportunity to learn the game.

In games Ia, II, III, and IV, two groups of five or six individuals each were recruited for each room; label these groups as A1, A2, B1, and B2. In the first session, each person in group A1 played a sequence of five one-shot games against different oppo-

nents from group B1, while group A2 played against group B2. In the second session, each person in group A1 played an additional sequence of five one-shot games against different opponents from group B2, while group A2 played against group B1. Game Ib was similar, except that there were three sessions, and three groups of seven individuals were recruited for each room. The rotation of groups across sessions was analogous to the other games: A1 played B1, then B2, then B3, etc.

At the beginning of each session, each participant was given a set of instructions (see Appendix A). The experimenter read the instructions aloud and answered questions. The size of the pie was determined at the start of each bargaining period by having the informed bargainer draw from a box containing N (50, 60, or 100) poker chips, including pN good (blue) chips and $(1-p)N$ bad (white) chips. Thus, it was presumably common knowledge that the probability of the good pie was p .

C. The Bargaining Experiment

In the bargaining games, the players negotiated over the amount of money that the informed bargainer would transfer to the uninformed bargainer; this transfer was not contingent on the informed bargainer's information. Each bargaining period lasted 10 minutes, during which free two-way communication was allowed by means of handwritten messages. The subjects gave messages to the monitor in their room, who passed them to the other bargainer via the monitor in the other room. A timekeeper was placed between rooms to mark the time on each message as it passed through and to ensure

that messages did not cross in the mail. Typically, a message arrived about 15 seconds after it was sent.

Each subject was provided with a large stock of forms on which to write messages. The forms included a space where formal proposals could be made and accepted. Once a formal proposal was made, it could be held by the other bargainer and accepted at any time before the end of the bargaining period. The first accepted proposal to reach the timekeeper determined the outcome for the period; if no proposal was accepted before the end of the period, each bargainer received zero for the period. The uninformed bargainer was never told the size of the pie that the informed bargainer had drawn, even when the period was over.

As can be seen from this description, the bargaining experiment was not faithful to the details of Myerson's model, since it did not explicitly involve bargaining over mechanisms. We take the view that the model is intended as an abstract approximation to realistic bargaining processes and ask whether this approximation will deliver useful predictions in a natural bargaining game.

Our experimental design also does not allow implementation of the full set of feasible mechanisms, but it does cover an interesting subset. The main restriction is that γ must be either 0 or 1 in the bad state. Efficiency requires in any case that there is no strike in the good state, so $\gamma_g = 1$. If the strike condition holds, then the mechanism determined by $\gamma_g = 1$, $\gamma_b = 0$, $x_g = \pi_g$, and $x_b = 0$ is efficient, and this would be optimal for the uninformed bargainer if he could dictate the mechanism. On the other hand, if a mechanism does not involve a strike, then incentive compatibility requires that the transfers x_b and x_g be equal. Therefore, our simplified design includes the most obvious mechanisms involving strikes, as well as all the relevant mechanisms in which there are no strikes.

D. The Random-Dictator Experiment

The RD games followed the same procedures, but instead of negotiating, each player submitted a single proposal that specified the amount the uninformed player should

receive. Then, the uninformed player drew from a box containing 50 red and 50 white chips to determine which player's proposal would dictate the outcome. If the selected proposal was not feasible (e.g., if π_b was drawn and the uninformed player demanded more than this), then neither player received anything. In tabulating the data, we used the proposals of both players, regardless of which player's proposal was randomly chosen.

III. Results: Strike Incidence

The analysis in Section I implies that strikes should occur only when the strike condition $p\pi_g \geq \pi_b$ holds and when π_b is drawn, if all players are selfish expected-income maximizers who bargain inscrutably over enforceable mechanisms. Thus, strikes can sometimes be construed as efficient when the size of the pie is not common knowledge. But if the incentive-efficiency principle is to serve as a general explanation for bargaining conflicts, the predictions made by this principle should not depend substantially on details of the bargaining process. In the bargaining experiment reported here, we examine a natural but complicated bargaining game to see if it may be reduced to a simple analysis of efficient mechanisms. Thus, we ask whether incentive efficiency helps explain the variations in strike incidence across the different states and parameter sets of the bargaining experiment.

We will show that incentive efficiency explains some, but not all, of the variation in strike incidence in the bargaining experiment. It might be that some of the unexplained variation occurred because the players were not selfish or not risk-neutral, as we assumed. We check this by comparing the bargaining results with the results of the RD games, in which any systematic discrepancy between predictions and outcomes can only be due to misspecification of the players' objectives. Alternatively, it might be that the free exchange of messages and proposals in the bargaining experiment introduced important strategic complications that were not captured in the abstract model of mechanism bargaining. In the next section,

TABLE 2—INCIDENCE OF STRIKES IN BARGAINING GAMES

Game	Pie size	Session 1	Session 2	Session 3	All sessions
Ia ($p = 0.5$)	\$1.00	48.4% (9.0%) (15/31)	29.6% (8.8%) (8/27)		39.7% (6.4%) (23/58)
	\$6.00	12.5% (6.9%) (3/24)	9.1% (6.3%) (2/22)		10.9% (4.6%) (5/46)
Ib ($p = 0.5$)	\$1.00	28.6% (6.0%) (16/56)	45.3% (6.8%) (24/53)	42.9% (6.6%) (24/56)	38.8% (3.8%) (64/165)
	\$6.00	10.2% (4.4%) (5/49)	3.9% (2.7%) (2/52)	10.4% (4.5%) (5/48)	8.1% (2.2%) (12/149)
I (Ia & Ib) ($p = 0.5$)	\$1.00	35.6% (5.2%) (31/87)	40.0% (5.5%) (32/80)	42.9% (6.6%) (24/56)	39.0% (3.3%) (87/223)
	\$6.00	11.0% (3.7%) (8/73)	5.4% (2.6%) (4/74)	10.4% (4.5%) (5/48)	8.7% (2.0%) (17/195)
II ($p = 0.75$)	\$2.30	21.4% (1.4%) (3/14)	13.3% (9.1%) (2/15)		17.2% (7.1%) (5/29)
	\$3.90	9.8% (4.7%) (4/41)	5.7% (4.0%) (2/35)		7.9% (3.1%) (6/76)
III ($p = 0.5$)	\$2.80	7.7% (5.3%) (2/26)	3.1% (3.1%) (1/32)		5.2% (2.9%) (3/58)
	\$4.20	5.9% (4.1%) (2/34)	7.1% (5.0%) (2/28)		6.5% (3.1%) (4/62)
IV ($p = 0.25$)	\$2.40	12.2% (5.2%) (5/41)	11.4% (4.8%) (5/44)		11.8% (3.5%) (10/85)
	\$6.80	5.3% (5.3%) (1/19)	0.0% (0.0%) (0/16)		2.9% (2.9%) (1/35)

Notes: The second line in each panel contains standard errors for the strike probabilities in the first line; the third line gives these probabilities as strikes/trials. The table excludes two cases in which an informed (but confused) player gave up more than the small pie even though the small pie was drawn.

TABLE 3—INCIDENCE OF STRIKES IN RANDOM-DICTATOR GAMES

Game	Pie size	Session 1	Session 2	All sessions
I ($p = 0.5$)	\$1.00	39.6% (7.1%) (19/48)	41.2% (6.0%) (28/68)	40.5% (4.6%) (47/116)
	\$6.00	0.0% (0/72)	0.0% (0/52)	0.0% (0/124)
II ($p = 0.75$)	\$2.30	17.6% (6.5%) (6/34)	27.3% (9.5%) (6/22)	21.4% (5.5%) (12/56)
	\$3.90	0.0% (0/76)	0.0% (0/78)	0.0% (0/154)
III ($p = 0.5$)	\$2.80	6.6% (3.2%) (4/61)	15.9% (5.5%) (7/44)	10.5% (3.0%) (11/105)
	\$4.20	0.0% (0/48)	0.0% (0/56)	0.0% (0/104)
IV ($p = 0.25$)	\$2.40	8.8% (3.4%) (6/68)	12.5% (3.7%) (10/80)	10.8% (2.6%) (16/148)
	\$6.80	0.0% (0/32)	0.0% (0/18)	0.0% (0/50)

Note: The second line in each panel referring to the bad pie contains standard errors for the strike probabilities in the first line; the third line gives these probabilities as (strikes/trials). The table excludes three cases in which an informed player gave up more than the small pie even though the small pie was drawn.

we will summarize our data on messages and proposals, with particular attention to possible violations of inscrutability.

Tables 2 and 3 show the incidence of strikes in the bargaining and RD experiments.⁶ It is apparent that the incentive-efficiency prediction does not work perfectly. In the bargaining experiment, strikes sometimes occurred even though the good pie was drawn, and in both experiments

strikes also occurred in the bad states of games III and IV, contrary to the efficiency prediction. These inefficient strikes would reject a strict version of the efficiency prediction, but such a rejection would be too harsh. In fact, there is generally a nonnegligible incidence of strikes in complete-information experiments, in which strikes cannot be construed as efficient.⁷ Our conclusion is

⁶We report results by session and also pooled over all sessions. To justify pooling, we tested and failed to reject the hypothesis that strike incidence was constant across sessions within each game. We also tested and failed to reject the hypothesis that our data were consistent with Nash equilibrium play. We did this by testing whether the IC constraints (1) were violated by the average player in each game (i.e., whether the informed player could gain by using the big-pie strategy when the pie was in fact small and vice versa).

⁷A review of complete-information bargaining experiments indicates that the incidence of apparently inefficient strikes in our experiments is not atypical. The following rough summary shows the range of strike-incidence results across the various treatments in each study (where a "strike" means failure to reach a Pareto optimum): 0–37% (Malouf and Roth, 1981); 5–35% (Janet Neelin et al., 1988); 8–33% (Roth and Murnighan, 1982); 10–29% (Jack Ochs and Roth, 1989); 10–29% (W. Guth et al., 1982); 19–42% (Binmore et al., 1985); 19–67% (Forsythe et al., 1990).

that the efficiency principle is successful in explaining the main fluctuations in strike incidence across games in our experiments: in both experiments, strike incidence was significantly higher when the small pie was drawn in games I and II.

Given that the good pie was drawn, the average strike incidence in the bargaining experiment was 7.6 percent, and the variation across games was not statistically significant. Adding in the bad states of games III and IV (to cover all cases in which efficient strikes are ruled out), the overall incidence was 8.0 percent (41 strikes in 511 trials), with no significant variation between cases. In the RD experiment, strikes never occurred in the good state, since no player ever demanded more than the good pie. However, when the bad pie was drawn, the incidence of strikes in each game was essentially the same as in the corresponding bargaining game.

Although incentive efficiency explains why strike incidence was high in the bad states of games I and II, it leaves substantial unexplained variation between these two games. In both the bargaining and the RD experiments, incidence in the bad state was significantly higher in games I and II than in games III and IV, but incidence was also significantly higher in game I than in game II. Notice that the informed player in game II could afford to concede half of the good pie (\$1.95) even if the bad pie (\$2.30) was actually drawn, while in game I the informed player could not afford to do this. In fact, we chose the parameters of games II and IV after observing the results of games I and III in the bargaining experiment, with the idea that strike incidence might be affected by whether the informed player could always afford to concede half of the good pie. This kind of concession was never made in the bargaining experiment, and in any case such concessions make no sense in the RD experiment, in which the results of games I and II were also significantly different. Appendix B shows the gap between the parties' best offers for each strike in the bargaining experiment. In many cases, the informed player was offering less than the small pie, while the uninformed player ei-

ther had made no offer or was demanding more than the small pie. Note, however, that all of the uninformed players' demands in games II and III could have been met, regardless of the state. In contrast, the informed players often could not afford to meet the uninformed players' demands in the bad state in games I and IV.

An alternative and more promising line of inquiry follows from the observation that the strike frequencies are remarkably similar across the bargaining and RD games. In the RD games, the unexplained variation in strike incidence can only be due to the misspecification of players' objectives. Suppose for example that some players were mildly altruistic or risk-averse. Such players might take \$1.15 for sure in game II, leaving \$1.15 or \$2.75 for the other player, while refusing to take \$0.50 in game I. This is discussed further in Section V below.

In summary, we have shown that the existence of unverifiable private information in a bargaining game may cause strikes even though free communication is allowed between the bargainers. We have also found support for the view that strikes are an efficient response to incentive-compatibility constraints, in the sense that strikes are much less likely to occur in situations where incentive-efficient strikes are not possible. In the next section, we will summarize the proposals and final agreements in the experiments and discuss evidence suggesting that at least some subjects were not selfish expected-income maximizers.

IV. Comparison of the Bargaining and Random-Dictator Games

We have shown that incentive efficiency gives a useful explanation of the strike-incidence results but that some anomalies remain. We now consider alternative explanations for the anomalies, concentrating on the question of why strike incidence was much higher in game I than in game II.

Table 4 summarizes the strike incidence results in Tables 2 and 3, showing the percentage of the pie that was lost in relation to how the pie was divided between the players (the columns labeled U and I show

TABLE 4—PIE SPLITS IN BARGAINING AND RANDOM-DICTATOR GAMES

Game	State	π	Payoffs		Value of information	Total payoff	Percentage loss
			U	I			
Bargaining games							
I ($p = 0.5$)	bad	1.00	0.31	0.30	-0.01	0.61	39.0
	good	6.00	1.78	3.70	1.92	5.48	8.7
	mean	3.50	1.05	2.00	0.95	3.05	13.0
	RD		1.50	1.75	0.25	3.25	7.1
II ($p = 0.75$)	bad	2.30	1.06	0.84	-0.21	1.90	17.2
	good	3.90	1.53	2.07	0.54	3.59	7.9
	mean	3.50	1.41	1.76	0.35	3.18	9.3
	RD		1.46	1.75	0.29	3.21	8.3
III ($p = 0.5$)	bad	2.80	1.47	1.18	-0.29	2.66	5.2
	good	4.20	1.52	2.41	0.89	3.93	6.5
	mean	3.50	1.50	1.80	0.30	3.29	6.0
	RD		1.40	2.10	0.70	3.50	0.0
IV ($p = 0.25$)	bad	2.40	1.08	1.04	-0.04	2.12	11.8
	good	6.80	1.58	5.03	3.45	6.61	2.9
	mean	3.50	1.21	2.04	0.83	3.24	7.4
	RD		1.20	2.30	1.10	3.50	0.0
Random-dictator games:							
I ($p = 0.5$)	bad	1.00	0.12	0.48	0.36	0.59	40.5
	good	6.00	2.57	3.43	0.86	6.00	0.0
	mean	3.50	1.35	1.96	0.61	3.30	5.7
	RD		1.50	1.75	0.25	3.25	7.1
II ($p = 0.75$)	bad	2.30	0.81	0.99	0.18	1.81	21.4
	good	3.90	1.49	2.41	0.92	3.90	0.0
	mean	3.50	1.32	2.06	0.74	3.38	3.6
	RD		1.46	1.75	0.29	3.21	8.3
III ($p = 0.5$)	bad	2.80	1.25	1.26	0.02	2.51	10.5
	good	4.20	1.61	2.59	0.98	4.20	0.0
	mean	3.50	1.43	1.93	0.50	3.36	4.1
	RD		1.40	2.10	0.70	3.50	0.0
IV ($p = 0.25$)	bad	2.40	0.97	1.17	0.20	2.14	10.8
	good	6.80	1.91	4.89	2.98	6.80	0.0
	mean	3.50	1.21	2.10	0.89	3.31	5.6
	RD		1.20	2.30	1.10	3.50	0.0

Note: There are minor inconsistencies across the columns of this table, because of rounding error.

the mean payoffs for the uninformed and informed players, counting strikes as zero payoffs). The table also shows the prediction of the RD axiom for each game. In the bargaining experiment, the uninformed player always did at least as well as the informed player in the bad state, while the informed player did better in the good state.

In the RD experiment, the informed player always did better than the uninformed player, regardless of pie size. In both experiments, the average value of the game was higher for the informed player. The percentage of the pie lost in the RD experiment was lower, on average, than in the corresponding games in the bargaining ex-

TABLE 5—BARGAINING GAMES VERSUS RANDOM-DICTATOR GAMES

Game	Bargaining games		Dictator games		<i>t</i> test		Wilcoxon test	
	N_1	Mean	N_2	Mean	<i>t</i>	<i>p</i>	<i>T</i>	<i>p</i>
Strikes:								
I, γ_b	223	0.61	116	0.59	-0.32	0.743	—	—
II, γ_b	29	0.83	56	0.79	-0.45	0.636	—	—
III, γ_b	58	0.95	106	0.90	-1.13	0.258	—	—
IV, γ_b	85	0.88	150	0.89	0.26	0.801		
Pie splits:								
I, x_b	223	\$0.31	116	\$0.12	-6.26	0.000	-5.35	0.000
y_b	223	\$0.30	116	\$0.48	4.50	0.000	2.61	0.008
II, x_b	29	\$1.06	56	\$0.81	-1.25	0.215	-1.38	0.176
y_b	29	\$0.84	56	\$0.99	0.75	0.458	0.15	0.883
III, x_b	58	\$1.47	105	\$1.25	-1.28	0.202	-1.07	0.286
y_b	58	\$1.18	105	\$1.26	0.45	0.655	0.13	0.912
IV, x_b	85	\$1.08	148	\$0.97	-0.83	0.410	-1.69	0.087
y_b	85	\$1.04	148	\$1.17	0.99	0.325	0.36	0.971
I, x_g	195	\$1.78	124	\$2.57	3.51	0.001	-0.25	0.786
y_g	195	\$3.70	124	\$3.43	-1.11	0.270	1.33	0.184
II, x_g	76	\$1.53	154	\$1.49	-0.20	0.838	1.21	0.231
y_g	76	\$2.07	154	\$2.41	2.03	0.044	0.22	0.823
III, x_g	62	\$1.52	104	\$1.61	0.47	0.641	0.91	0.377
y_g	62	\$2.41	104	\$2.59	0.87	0.387	0.43	0.668
IV, x_g	35	\$1.58	50	\$1.91	1.06	0.317	0.94	0.351
y_g	35	\$5.03	50	\$4.89	-0.40	0.687	-0.65	0.517

Note: The Wilcoxon signed-ranks statistic D is the sum of the ranks of the random-dictator observations. If N_1 is the number of observations in the bargaining game and N_2 is the number of observations with the random-dictator game, the T value we report is⁵

$$T = (D - N_1 N_2 / 2) / [(N_1 N_2 (N_1 + N_2 + 1) / 12)]^{1/2}$$

which is approximately a standard normal random variable for large values of N_1 and N_2 . This table excludes five cases in which the informed player gave up more than the small pie even though the small pie was drawn.

periment, mainly because strikes did not occur in the good state of the RD experiment.

The main difference in design between the two experiments is that the RD experiment suppresses the communication of messages and proposals; in addition, each player in the bargaining experiment had the right to reject any demand made by the other. On the face of it, suppressing communication makes a big difference, since there is an enormous variety of messages which could be sent, and it would be impossible to write down a complete strategy for the bargaining game, specifying the message to be sent after each possible history. From another point of view, the communication in the

bargaining experiment is merely cheap talk, since the informed player can send no message that credibly reveals that the pie is small, and if both players are selfish expected-income maximizers, this is the only piece of information that is relevant. This cheap-talk interpretation is based on stringent assumptions about preferences, however; what if some of the subjects find it difficult to lie? Moreover, even if the preference assumptions are valid, Joseph Farrell and Robert Gibbons (1989) have shown that cheap talk can support bargaining equilibria that would not otherwise be possible.

Our results show that the general pattern of the outcomes in the bargaining games was very similar to that of the RD games.

Communication did not substantially affect the incidence of strikes in the bad state; however, in the good state, strikes occurred only in the bargaining games, where the informed player could insist that the pie was small and the uninformed player had the right to insist that it was not. Moreover, Table 5 shows that, with the exception of game I, there were no statistically significant differences in pie splits between the two experiments. Furthermore, the unexplained variation in strike activity between games I and II occurred in both experiments, so the source of this variation apparently lies in the subjects' preferences, rather than in any messages that passed between them.

A. Altruism

In the bargaining games, violations of the inscrutability assumption might be interpreted as a concern for fairness, because revealing private information benefits both players collectively by reducing strikes. More generally, offering more than is necessary to get the other player to agree to one's demand would be evidence of altruism. If informed players behaved inscrutably in the bargaining games, then the conditional distribution of offers, given that the big pie is drawn, should be the same as the conditional distribution given the small pie. At the other extreme, if an informed player offers more than the small pie, the uninformed player can confidently infer that the big pie has been drawn, and the game transposes to one of complete information; we call this a "revealing offer."

Table 6 shows the frequency distributions of "best offers" made by the informed players, conditioned on whether the pie was big or small. The informed players behaved inscrutably in some, but not all, of these games. In game III, for instance, the conditional distributions are almost identical: the informed player usually offered half of the small pie (\$1.40) and sometimes offered a little more, regardless of whether the pie was in fact big or small. In game II, on the other hand, the informed player was much more likely to make a favorable offer in the

good state. Although the data are less clear, it seems reasonable to infer that inscrutability failed in games I and IV also.⁸

Revealing offers were made surprisingly often in game I, as can be seen in Table 6. In games II and III, a revealing offer requires that the informed player give up substantially more than half of the big pie; as one might expect, this never happened. In game IV, as in game I, the big pie (\$6.80) was more than twice the small pie (\$2.40), yet revealing offers were made in only 2 of 35 trials in game IV. A summary of revealing offers is given in Table 7. The chi-square tests shown there support the hypothesis that these offers were just as likely in the second experimental session as in the first.

We cannot conclude that violations of inscrutability result from players' concerns for fairness in the bargaining games, since we do not know the minimum offer necessary to achieve an agreement in these games. In the RD games, however, players could unilaterally impose the outcome. Thus, informed players could simply give their opponents nothing, and uninformed players should always take at least the small pie. Deviations from these strategies can only be due to altruism. In 18.6 percent (80/430) of the RD games, the informed player gave the uninformed player more than ten cents.⁹ This happened less often in session II (7.1 percent, or 15/210) than in session I (29.5 percent, or 65/220). In session I, 12.7 percent (28/220) of the uninformed players demanded less than the bad pie, and the corresponding figure in session II was 4.3 percent (9/210). We conclude that the typical player in these games behaved selfishly but that altruism, although not typical, had a substantial influence on the outcomes.

⁸Of course, each player saw only a fragment of Table 6, so that even in those games where inscrutability failed, the informed player's behavior might not convey much information to the uninformed player.

⁹Some informed players gave their opponents a penny (eight instances), a nickel (two instances), or a dime (17 instances). We prefer not to consider such trifling amounts as evidence of altruism.

TABLE 6—BEST OFFERS BY INFORMED PLAYERS: DOES INSCRUTABILITY HOLD?

Game Ia:

Pie = \$1.00			Pie = \$6.00		
Offer	Frequency	Percentage	Offer	Frequency	Percentage
0.40	2	3.39	0.50	25	54.35
0.45	1	1.69			
0.50	46	77.97	0.51	1	2.17
0.52	1	1.69	0.55	2	4.35
0.55	4	6.78	0.65	2	4.35
0.60	3	5.08	1.50	1	2.17
0.80	1	1.69	2.00	2	4.35
0.96	1	1.69	2.20	1	2.17
Total:	59	100.00	2.25	1	2.17
			2.50	1	2.17
			2.75	2	4.35
			2.90	1	2.17
			2.96	1	2.17
			3.00	6	13.04
			Total:	46	100.00

Game Ib:

Pie = \$1.00			Pie = \$6.00		
Offer	Frequency	Percentage	Offer	Frequency	Percentage
0.45	5	3.03	0.50	65	43.92
0.47	1	0.61			
0.50	143	86.67	0.51	2	1.35
0.51	1	0.61	0.52	1	0.68
0.53	1	0.61	0.54	1	0.68
0.55	6	3.64	0.55	3	2.03
0.57	1	0.61	0.60	9	6.08
0.60	2	1.21	0.65	2	1.35
0.67	1	0.61	0.70	4	2.70
0.75	3	1.82	0.75	11	7.43
1.00	1	0.61	0.80	2	1.35
Total:	165	100.00	0.95	1	0.68
			1.00	7	4.73
			1.50	4	2.70
			2.00	15	8.78
			2.25	2	1.35
			2.50	4	2.70
			2.55	1	0.68
			2.70	1	0.68
			2.75	2	1.35
			3.00	13	8.78
			Total:	148	100.00

Game II:

Pie = \$2.30			Pie = \$3.90		
Offer	Frequency	Percentage	Offer	Frequency	Percentage
1.10	1	3.45	1.00	2	2.70
1.15	16	55.17	1.15	21	28.38
1.20	2	6.90	1.20	2	2.70
1.25	1	3.45	1.25	3	4.05
1.30	2	6.90	1.30	10	13.51
1.35	2	6.90	1.38	1	1.35
1.46	1	3.45	1.50	5	6.76
1.50	1	3.45	1.55	1	1.35
1.55	1	3.45	1.60	1	1.35
1.70	1	3.45	1.65	1	1.35
1.75	1	3.45	1.68	1	1.35
Total:	29	100.00	1.75	7	9.46
			1.80	3	4.05
			1.85	1	1.35
			1.90	3	4.05
			1.95	12	16.22
			Total:	74	100.00

TABLE 6—Continued

Game III:			Pie = \$4.20		
Pie = \$2.80			Offer	Frequency	Percentage
Offer	Frequency	Percentage	1.20	1	1.61
1.00	1	1.72	1.30	1	1.61
1.20	1	1.72	1.35	2	3.23
1.35	2	3.45	1.37	1	1.61
1.40	36	62.07	1.40	42	67.74
1.45	3	5.17	1.41	1	1.61
1.50	4	6.90	1.45	1	1.61
1.60	1	1.72	1.50	1	1.61
1.70	3	5.17	1.60	4	6.45
1.75	2	3.45	1.75	2	3.23
1.80	4	6.90	1.80	1	1.61
2.10	1	1.72	1.90	1	1.61
Total:	58	100.00	2.00	2	3.23
			2.10	2	3.23
			Total:	62	100.00

Game IV:			Pie = \$6.80		
Pie = \$2.40			Offer	Frequency	Percentage
Offer	Frequency	Percentage	1.15	1	2.86
0.80	2	2.35	1.20	22	62.86
0.85	1	1.18	1.30	1	2.86
0.95	1	1.18	1.40	1	2.86
1.00	8	9.41	1.50	6	17.14
1.10	2	2.35	1.60	1	2.86
1.15	2	2.35	2.00	1	2.86
1.16	1	1.18	3.00	1	2.86
1.19	2	2.35	3.40	1	2.86
1.20	57	67.06	Total:	35	100.00
1.21	1	1.18			
1.25	1	1.18			
1.30	2	2.35			
1.40	2	2.35			
1.50	2	2.35			
2.40	1	1.18			
Total:	85	100.00			

TABLE 7—REVEALING OFFERS

Reveal?	Game Ia			Game Ib				Game IV
	Session		Total	Session			Total	Total
	1	2		1	2	3		
No	17	12	29	36	34	31	101	33
Yes	7	10	17	13	18	17	48	2
Total:	24	22	46	49	52	48	149	35
$X^2_{[1]} = 1.3070, p = 0.25$ $X^2_{[2]} = 0.9727, p = 0.62$								

Note: A revealing offer was defined as an offer made by the informed player of at least the size of the small pie.

TABLE 8—UNINFORMED PROPOSALS IN RANDOM-DICTATOR GAMES

Game I				Game II			
Uninformed proposal	Session		Total	Uninformed proposal	Session		Total
	1	2			1	2	
1.00	14	8	22 (18%)	1.15	1	1	2
3.00	1	0	1	1.50	1	0	1
4.00	1	0	1	1.95	6	2	8
5.00	8	0	8	2.00	7	0	7
5.50	0	1	1	2.15	1	0	1
5.99	1	0	1	2.25	1	0	1
6.00	35	51	86 (72%)	2.30	23	28	51 (49%)
				3.00	3	0	3
Total:	60	60	120	3.90	12	19	31 (30%)
				Total:	55	50	105

Game III				Game IV			
Uninformed proposal	Session		Total	Uninformed proposal	Session		Total
	1	2			1	2	
0.50	2	1	3	0.00	1	0	1
1.00	3	0	3	0.75	0	1	1
2.00	1	0	1	1.40	1	0	1
2.20	1	0	1	2.00	1	1	2
2.40	0	1	1	2.30	1	0	1
2.75	0	2	2	2.40	37	36	73 (73%)
2.80	41	34	75 (71%)	2.80	1	0	1
3.00	1	0	1	3.00	2	1	3
4.20	6	12	18 (17%)	4.00	2	1	3
Total:	55	50	105	6.80	4	10	14 (14%)
				Total:	50	50	100

B. Attitudes Toward Risk

Table 8 lists the demands of the uninformed players in each game of the RD experiment. Selfish and risk-neutral players would demand π_g in games I and II, and π_b in games III and IV. The typical player behaved this way in games I, III, and IV, but not in game II. The expected value of the tough demand in game II was \$2.93, but about half of the players instead chose \$2.30 as a sure thing. This might be due to mild risk aversion combined with some concern for fairness (since the tough demand leaves nothing for the informed player, while the weak demand leaves \$1.60 when the pie is big). The typical player in game I made the tough demand, but a substantial minority (18 percent) appeared to be strongly risk-averse, choosing \$1.00 for sure over a 50-percent chance of \$6.00. At the other ex-

treme, a substantial minority (21 percent) of the players in game IV appeared to be risk-seekers, choosing a 25-percent chance of getting (at most) \$6.80 over \$2.40 as a sure thing. We conclude that while the typical subject appeared to be risk-neutral, there were important deviations from this type in both directions.¹⁰

¹⁰If the subjects were all selfish expected-utility maximizers (and if this were common knowledge), the heterogeneity in risk attitudes could be eliminated by using the ingenious binary lottery procedure introduced by Roth and Malouf (1979). We chose not to use this procedure, mainly because it introduces considerable additional complexity in an already complicated experimental environment. In addition, the procedure works only under assumptions which are implausible in our context: that each subject acts selfishly and obeys the compound lottery axiom and that this is common knowledge.

C. Heterogeneity of Preferences

As Table 8 shows, the RD experiment uncovered considerable heterogeneity in subjects' preferences with regard to fairness and risk. This heterogeneity may help explain the variations in strike incidence in the bargaining experiment. In particular, although the uninformed player might maximize the expected payoff by playing tough in both games I and II, the expected loss from playing soft was less in game II, so that this option was more attractive to some risk-averse or altruistic players. On the other hand, although the soft strategy might maximize expected payoff in game IV, the tough strategy was apparently more attractive to a minority of risk-loving players. Since we have repeated observations on each individual, we can measure the extent to which atypical outcomes in each game are due to the presence of subjects with atypical preferences.¹¹

Table 9 shows individual effects on strike incidence. First consider the RD experiment, which directly measures heterogeneity. Table 9A shows the number of times each uninformed player proposed a strike (i.e., made a proposal that exceeded the small pie). In games I and II, a risk-neutral and selfish player should always make such a proposal, and in games III and IV, never. The results show that there are indeed some consistent differences in behavior across subjects. In game I, 16 of 24 players issued strike proposals in all five games in which they were uninformed, while only one of the 24 consistently chose to avoid a strike. A surprising result is that the other seven players in game I behaved inconsistently, sometimes choosing a strike proposal and sometimes not (these inconsistencies persisted in the second session, after the play-

ers had had plenty of time to learn the game). One might infer that these players were sufficiently altruistic or sufficiently risk-averse that they were indifferent between tough and soft proposals. By contrast with game I, only three of 20 players consistently chose to strike in game II, while nine of the 20 players behaved inconsistently. In games III and IV, only one player consistently chose to strike, and there was a slight majority of players who consistently played soft in each game.

In the bargaining games, the measurement of individual effects is less clear-cut, since the number of times each subject played against an informed player with a small pie was determined by chance. Table 9B shows the distribution of strikes over individuals, given that the small pie was drawn. Consistency of the outcomes in the bargaining games means either an entry in the first column (where the uninformed player is never involved in a strike) or else an entry on the main diagonal (where the uninformed player is involved in a strike every time the small pie is drawn). In game Ib, for instance, six of the 42 players were always involved in strikes, 15 were never involved, and the other 21 were sometimes involved in strikes. The small gap between strike probabilities in games II and IV, which was shown in Table 2, virtually disappears in Table 9: here 76 percent (13/17) of the uninformed players in game II were never involved in (small-pie) strikes, while the corresponding figure for game IV was 75 percent (18/24). On the other hand, the prevalence of strikes in game I stands out even more clearly here: only 30 percent (19/62) of the uninformed players were never involved in strikes.

Thus, it appears that there was a substantial degree of heterogeneity across players which blunted the incentive-efficiency prediction that strikes should occur only when the bad pie is drawn and when a risk-neutral and selfish player would rationally choose to play tough. However, a fair summary of the results is that the incentive-efficiency prediction works well when the expected payoff from being tough is large

¹¹A formal statistical analysis of individual effects would be possible, but only if the experiment were greatly enlarged. For example, we have only five measurements from which to infer the idiosyncratic behavior of each subject as an uninformed player, and these same measurements must also be used to draw inferences about the informed players on the other side.

TABLE 9—INDIVIDUAL STRIKE EFFECTS IN BARGAINING AND RANDOM-DICTATOR GAMES

A. Uninformed Players in Random-Dictator Games^a

Game	Number of Strikes						Total
	0	1	2	3	4	5	
I	1	2	1	2	2	16	24
II	8	3	4	0	2	3	20
III	12	4	1	0	2	1	20
IV	10	4	3	1	2	0	20

B. Uninformed Players in Bargaining Games

Game Ia						Game Ib								
Number of small pies	Number of strikes					Number of small pies	Number of strikes							
	0	1	2	4	Total		0	1	2	3	4	5	6	Total
1	2	0	—	—	2	1	2	1	—	—	—	—	—	3
2	1	3	1	—	5	2	2	0	2	—	—	—	—	4
3	1	4	1	—	6	3	3	0	2	1	—	—	—	6
4	0	4	1	1	6	4	5	2	2	2	1	—	—	12
5	0	0	1	0	1	5	3	3	6	1	0	1	—	14
Total:	4	11	4	1	20	6	0	0	1	0	0	1	0	2
						7	0	0	0	0	0	0	1	1
						Total:	15	6	15	4	1	2	1	42

Game II					Game III				Game IV					
Number of small pies	Number of strikes				Number of small pies	Number of strikes			Number of small pies	Number of strikes				Total
	0	1	2	Total		0	1	Total		0	1	2	3	
1	7	1	—	8	1	4	0	4	1	1	0	—	—	1
2	4	2	1	7	2	4	1	5	2	3	0	0	—	3
3	1	0	0	1	3	6	2	8	3	7	0	0	0	7
4	1	0	0	1	4	5	0	5	4	6	1	0	1	8
Total:	13	3	1	17	Total:	19	3	22	5	1	2	2	0	5
									Total:	18	3	2	1	24

Note: A strike means that the uninformed player demanded more than the small pie.

^aThe table excludes two players in game II, and two in game III, because they did not participate in all 10 periods.

enough (relative to the payoff from being soft) to overcome a mild dose of risk aversion or altruism.¹²

¹²One might reduce the problem of heterogeneity by conducting the experiment in two phases, using the first phase to screen out undesirable subjects; but if, for example, no two subjects are exactly alike in the extent to which they are altruistic, risk-averse, or consistent, the heterogeneity problem cannot be eliminated. Moreover, there are additional strategic complications in the screening phase of the game, since some players who should be screened out might dissemble if they desire to play the second phase of the game. This problem could be reduced by making the second phase

V. Conclusions

We have carried out an experiment designed to analyze the effect of one-sided private information on the outcomes of two-person bargaining games. The main question is whether "strikes" (failure to reach agreement) can be interpreted as efficient responses to the incentive constraints arising from private information. The bar-

attractive only to the subjects who are targeted for study, but the analysis of the entire game is likely to be very complicated.

gaining experiment allowed free exchange of messages between the two players, which makes the games more realistic, but which also means that a strategic analysis would be impossibly complicated. We therefore considered an abstract approximation of these games, in which the players are supposed to bargain over alternative mechanisms which might be used to divide the pie. For some parameter values, all mechanisms involving strikes are dominated, according to the criterion of incentive-efficiency, and thus strikes should not occur. Our experiment included two games of this kind, as well as two games in which efficient strikes could occur in the bad state.

The main conclusion is that the abstract principle of incentive-efficiency is very helpful in explaining the actual incidence of strikes in our experimental bargaining games. In general, strikes occurred much more often in the bad states of games I and II, where the condition for efficient strikes was satisfied, than in those situations where strikes were inefficient. Since our experiment was designed as a compromise between realistic bargaining games and Myerson's (1984) abstract model of bargaining over mechanisms, the results suggest that the incentive-efficiency principle may prove robust in applications.

The incentive-efficiency principle identifies circumstances in which strikes should not occur, but when efficient strikes are possible it does not predict how often they will actually happen. We used Myerson's random-dictator axiom to obtain more specific predictions, but these predictions performed poorly in some cases with regard to both the incidence of strikes and the division of the pie. The random-dictator axiom asserts that the outcome will be as if the two parties randomized equally between the mechanisms each would unilaterally impose, if given dictatorial power. This implies that strikes should occur 50 percent of the time in the bad states of games I and II; but in fact, strikes were much more frequent in game I than in game II.

We carried out another experiment designed to test the random-dictator axiom directly, with the idea that perhaps the fail-

ure of this axiom was due to the use of unconstrained "cheap talk" in our bargaining experiment. In fact, however, the random-dictator results were remarkably similar to those of the bargaining games. We concluded that the random-dictator axiom failed because there was considerable heterogeneity among our subjects, including sizeable minorities of both risk-averse and risk-loving types, and another minority of altruists.

Although we have some encouraging results on the predictive power of the incentive-efficiency principle, the results on heterogeneity of preferences and on inconsistency of decisions indicate that much caution is needed in drawing conclusions from behavior in bargaining (and other) experiments. In any game in which the players interact strategically, the theoretical analysis should not begin (as ours did) with the assumption that the players' objective functions are common knowledge. In addition, given that there is a nontrivial distribution of objective functions in the population of potential subjects, it is difficult to see how the actual subjects could know what this distribution is.

APPENDIX A

Instructions for Bargaining Game [Exact Transcript]

General

You have been asked to participate in an experiment, consisting of two sessions, in the economics of bargaining. If you follow these instructions carefully and make good decisions, you might earn a considerable amount of money, which will be paid to you in cash at the end of the second experimental session.

Each experimental session will consist of five separate bargaining periods lasting ten minutes each. In each of these periods you will be bargaining with a different person who is in another room. Once this ten minute period is over, you will never bargain with this person again. The individuals you bargain with in the first session are different from the individuals you bargain with in the second session. You will not be told who these people are either during or after the experiment. In each of these ten bargaining periods, you will negotiate with one of these persons to decide how to divide a sum of money.

You will notice that there are other people in the same room with you who are also participating in bargaining experiments. You will never be bargaining with any of these people during either bargaining ses-

sion. The decisions that they make will have absolutely no effect on you nor will any of your decisions affect them.

In each period, one bargainer will have more information than the other. In particular, the bargainer in one room will know the amount of money that is available to be bargained over for the period. The bargainer in the other room will only know that it is one of two possible amounts, \$6.80 or \$2.40. We will alternate which set of bargainers will be INFORMED and which will be UNINFORMED from one period to the next. In particular, if you are in room A you will be informed during all even numbered periods while if you are in room B you will be informed during all odd numbered periods. Thus, you will alternate between being an INFORMED bargainer and an UNINFORMED bargainer during the ten bargaining periods of the experiment.

The amount of money you will be bargaining over in a bargaining period will be determined at the start of each period. In each period, the \$2.40 amount is three times as likely to occur as the \$6.80 amount, regardless of what has happened previously. Before a bargaining period begins, each INFORMED bargainer will draw from a box containing one hundred poker chips. Seventy-five of the chips are white and correspond to the \$2.40 amount while the other twenty-five chips are blue and correspond to the \$6.80 amount. Thus, if a white chip is drawn, you will be bargaining over \$2.40 while if a blue chip is drawn you will be bargaining over \$6.80. At the beginning of each period, a separate, independent drawing will be held which determines the amount to be bargained over for that period. The amount to be bargained over in a given period will depend only on the color of the chip that is drawn for that particular period. In particular, a white chip is three times more likely to be drawn than a blue chip at the beginning of any period. The outcome of the drawing in any one period will have no effect on the outcome of the drawing in any other period.

A bargaining period will last for 10 minutes. During this time you will negotiate over how the money is to be divided. The bargaining will be over the amount the UNINFORMED bargainer is to receive. The amount of money that the INFORMED bargainer receives will be the actual amount of money at stake minus the amount that the UNINFORMED bargainer receives. Neither bargainer earns anything for the period unless an agreement is reached.

In your folder you will find two profit sheets—one for each session in which you will participate. Your “Beginning Cash Balance” is shown on the first row of your profit sheets. If this is your first session, this amount is \$10. If this is your second session, this is your cash balance from the end of the first session. The information next to each bargaining period number tells you whether you are the INFORMED or the UNINFORMED bargainer for that period. At the beginning of each period, the INFORMED bargainer should fill in the “Amount Bargained Over” for that period. Recall that this amount is determined by the color of the chip that the INFORMED bargainer draws from the box. If you and the other bargainer reach an agreement in a period, you should fill in the amount you receive on the row labeled “Your Earnings” for

that period. If you are the UNINFORMED bargainer, this is the amount you agreed to. If you are the INFORMED bargainer, this is the “Amount Bargained Over” minus the amount you agreed that the UNINFORMED bargainer should receive. If you and the other bargainer did not reach an agreement during the ten-minute bargaining period, you should record zero as your earnings for that period.

At the end of an experimental session, add “Your Earnings” from each of the five periods to your “Beginning Cash Balance” and record the total in the row corresponding to the “End of Session Cash Balance.” If this is the end of your first experimental session, this amount will be your “Beginning Cash Balance” when you begin the second session. If this is the end of your second experimental session, the experimenter will pay you this amount in cash.

Bargaining Rules

You will be able to bargain by sending messages and formal proposals to the other bargainer. On your table you will find an ample supply of a form labeled “Message Form.” You must fill in the number of the other bargainer and the bargaining period on this form. In your messages, you may communicate whatever you wish, except that you cannot make physical threats or reveal your identity. If you wish to send a message simply hand it to the experimenter who will deliver it for you, unless there is a message from the other bargainer about to be delivered to you. (That is, messages will not be allowed to cross in the mail.)

If you wish to make a formal proposal fill in an amount on the bottom line of your message form where it states: “I propose that the uninformed bargainer receive.” Once you have made a formal proposal, the other bargainer can accept it at any time; you can not take it back. If you wish to accept a proposal which has been made by the other bargainer, you must check the space marked “accept” on this proposal form and hand it to the experimenter.

You may send as many messages as you wish. You may also send as many formal proposals as you wish but remember that sending a new proposal does not cancel any of your previous proposals. You may have as many formal proposals outstanding during a bargaining period as you wish. The person you are bargaining with is free to accept at most one of these proposals. Similarly, you may receive many formal proposals from the person you are bargaining with and you may accept at most one of these proposals.

If you accept a formal proposal that was made by the other bargainer or if the other bargainer accepts a proposal which you made, this acceptance constitutes a binding contract and determines your earnings for that period. If no proposal has been accepted when the buzzer sounds to indicate the end of the period, neither bargainer receives anything for that period.

If you accept a proposal during a bargaining period, you should record your earnings from the agreement on your Profit Sheet at that time and wait for the next bargaining period to begin. You will need to change this amount only if the other bargainer accepts one of your proposals first. If no proposal is accepted in a bargaining period before the time limit is exceeded,

you should record "0" as your earnings on your Profit Sheet.

You will keep track of the time remaining in any period yourself, using the clock in your room or your own watch. When you receive a message, there will be a number in the upper right hand corner of the message form. This indicates the amount of time left in the current bargaining period. We will only tell you when a period is starting and when it is ending. We ask that you do not talk to other bargainers in your room until your bargaining session is completed. Do not be concerned if other bargainers finish a particular bargaining period before you complete it. A new period will not begin until all bargainers have completed bargaining for the current period.

In summary, your earnings in the experiment will be the total of your beginning cash balance plus your share of any agreements reached during the experiment. The amount of money you earn will depend partly upon luck and partly upon whether you have made good decisions. Your earnings will be exactly what you agreed to in the course of bargaining, as indicated on your profit sheets. Are there any questions?

Instructions for Random-Dictator Game
[Exact Transcript]

General

You have been asked to participate in an economics experiment consisting of two sessions. If you follow these instructions carefully and make good decisions, you might earn a considerable amount of money, which will be paid to you in cash at the end of the second experimental session.

Each experimental session will consist of five separate periods. In each of these periods you will be paired with a different person who is in another room. Once this period is over, you will never be paired with this person again. The individuals you are paired with in the first session are different from the individuals you are paired with in the second session. You will not be told who these people are either during or after the experiment. In each of these ten periods, you will be paired with one of these persons to decide how to divide a sum of money.

You will notice that there are other people in the same room with you who are also participating in the experiment. You will never be paired with any of these people during either session. The decisions that they make will have absolutely no effect on you nor will any of your decisions affect them.

In each period, one person in each pair will have more information than the other. In particular, the person in one room will know the amount of money that is available for the period. The person in the other room will only know that it is one of two possible amounts, \$6.00 or \$1.00. We will alternate which set of people will be INFORMED and which will be UNINFORMED from one period to the next. In particular, if you are in room A you will be informed during all even numbered periods while if you are in room B you will be informed during all odd numbered periods. Thus, you will alternate between being INFORMED

and UNINFORMED during the ten periods of the experiment.

Determining the Amount Of Money Available in Each Period

The amount of money available to be divided between each pair will be determined at the start of each period. In each period, the \$1.00 amount is just as likely to occur as the \$6.00 amount, regardless of what has happened previously. Before a period begins, each INFORMED person will draw a chip from a box labeled "MONEY BOX." This box contains one hundred poker chips. Fifty of the chips are white and correspond to the \$1.00 amount while the other fifty chips are blue and correspond to the \$6.00 amount. Thus, if a white chip is drawn the amount available will be \$1.00 while if a blue chip is drawn the amount available will be \$6.00. At the beginning of each period, a separate, independent drawing will be held which determines the amount available for that period. The amount available in a given period will depend only on the color of the chip that is drawn for that particular period. In particular, a white chip is just as likely to be drawn as a blue chip at the beginning of any period. The outcome of the drawing in any one period will have no effect on the outcome of the drawing in any other period.

Determining the Amount Each Person Receives

In a period you must propose how the money is to be divided. Each person will have to make a proposal which specifies the amount the UNINFORMED person is to receive. To do this, you must submit a proposal on a "Proposal Form." You will find an ample supply of these forms on the table in front of you. You must fill in the period number and the number of the other person you are paired with on this form. Once you have come to a decision about how much you wish to propose that the UNINFORMED person should receive, simply write that amount on a proposal form and hand it to the experimenter.

After all people have submitted their proposals, the UNINFORMED person will draw a chip from a box labeled "Proposal Box". This box contains one hundred chips. Fifty of the chips are red while the other fifty of the chips are white. The color of the chip which is drawn will determine whether the INFORMED person's proposal or the UNINFORMED person's proposal is used. If a red chip is drawn, then the UNINFORMED person's proposal will be used, while if a white chip is drawn then the INFORMED person's proposal will be used. Remember that the box contains an equal number of white chips and red chips, so that each person's proposal will have an equal chance of being selected.

The amount of money which each person receives is determined as follows:

- 1) If the selected proposal specifies that the UNINFORMED person should receive more than the amount of money available, neither person will earn anything for the period.
- 2) If the selected proposal specifies that the UNINFORMED person should receive an amount which

does not exceed the amount of money available, then the UNINFORMED person will receive the amount of money specified in the proposal. The INFORMED person will receive the amount of money which is left over.

Do not be concerned if other people finish a particular period before you complete it. A new period will not begin until all people have completed the current period.

If you have any questions during the experiment, ask the monitor in your room and he or she will answer them for you. Other than these questions, you *MUST* keep silent until your session is completed. If you break this silence while the experiment is in progress, you will be given one warning. If you continue to talk after you have been warned, you will be asked to leave the experiment and YOU WILL FORFEIT YOUR EARNINGS.

Recording Rules

In your folder you will find an information sheet. Your "Beginning Cash Balance" of \$10 is shown on the first row of your profit sheet. The information next to each period number tells you whether you are the INFORMED or the UNINFORMED person for that period. It also tells you the number of the person you are paired with for the period. During those periods in which you are the INFORMED person you should fill in the "Amount of Money Available" for that period. Recall that this amount is determined by the color of the chip that the INFORMED person draws from the "Money Box." During those periods in which you are the UNINFORMED person you should circle whether the informed person's proposal or the uninformed person's proposal has been selected. Recall that the proposal that will be used is determined by the color of

the chip that the UNINFORMED person draws from the "Proposal Box."

At the end of a period, we will give you a copy of your profit sheet for the period. We have included a blank profit sheet in your folder which you should look at now. On this sheet we will have recorded the amount of your proposal on the line labeled "Your Proposal." We will also have recorded the other person's proposal on the line labeled "Other Proposal." The proposal which we have circled is the one that will be used in this period. Finally, we have recorded your profits for the period on the line labeled "You Receive." Recall that this amount will be zero if the selected proposal exceeds the amount of money available. Otherwise, this is the amount specified on the selected proposal if you are the UNINFORMED person. If you are the INFORMED person, this is the "amount of money available" minus the amount that the UNINFORMED person receives. You should record your profits for the period on your information sheet.

At the end of an experimental session, add the amount you have received from each of the five periods to your "Beginning Cash Balance" and record the total in the row corresponding to the "End of Session Cash Balance." If this is the end of your first experimental session, this amount will be your "Beginning Cash Balance" when you begin the second session. If this is the end of your second experimental session, the experimenter will pay you this amount in cash.

In summary, your earnings in the experiment will be the total of your beginning cash balance plus the amount you earn during each period of the experiment. The amount of money you earn will depend partly upon luck and partly upon whether you have made good decisions. Your earnings will be exactly what is indicated on your profit sheets. Are there any questions?

APPENDIX B

Best Offers Prior to Strikes

Game	Pie size	Informed		Uninformed		Player ID	
		Time	Amount	Amount	Time	Uninformed	Informed
Ia	big (\$6.00)	9:31	2.50	3.00	8:17	13	8
		0:41	1.50	3.50	2:25	15	14
		9:52	0.50	—	—	17	22
		0:38	2.00	3.00	9:05	18	7
		0:26	0.50	—	—	21	22
	small (\$1.00)	1:02	0.50	3.50	2:18	1	16
		8:53	0.50	3.00	2:06	1	20
		9:54	0.50	—	—	3	12
		1:13	0.50	—	—	4	15
		4:40	0.50	—	—	5	6
		6:08	0.50	2.50	3:49	9	2
		9:08	0.50	3.00	0:16	9	6
		9:48	0.50	2.70	6:28	10	3
		9:54	0.50	—	—	11	4
		3:42	0.50	—	—	11	12

Best Offers Prior to Strikes—Continued

Game	Pie size	Informed		Uninformed		Player ID	
		Time	Amount	Amount	Time	Uninformed	Informed
		1:46	0.40	—	—	13	14
		8:50	0.50	2.00	1:47	14	1
		8:48	0.50	1.95	0:53	14	9
		1:36	0.60	2.10	0:42	14	17
		3:51	0.45	2.90	0:10	14	21
		8:41	0.50	3.00	2:51	15	6
		0:52	0.50	3.00	0:37	16	13
		1:10	0.50	—	—	17	16
		9:51	0.50	0.90	0:58	19	10
		2:21	0.96	2.70	0:20	21	20
		0:22	0.40	3.00	4:43	22	15
		8:92	0.50	3.00	0:10	22	19
Ib	big (\$6.00)	0:50	0.50	1.75	0:20	5	10
		0:56	1.50	2.90	3:00	5	34
		9:46	0.50	1.00	0:00	12	9
		9:33	0.50	2.50	2:40	16	1
		9:58	0.50	2.50	1:28	18	25
		0:00	0.54	—	—	19	20
		0:04	3.00	3.00	2:00	25	6
		0:38	2.00	3.00	1:39	27	14
		1:06	2.50	4.00	1:28	28	21
		1:48	0.75	3.50	2:32	33	24
		3:17	0.75	3.00	9:00	40	9
		9:41	0.50	2.50	1:28	42	27
Ib	small (\$1.00)	6:06	0.50	—	—	1	10
		7:21	0.50	3.00	8:11	1	22
		8:54	0.50	3.00	0:15	1	34
		0:40	0.75	2.00	1:02	3	24
		2:08	1.00	4.00	1:47	4	25
		0:11	0.50	3.50	0:18	4	37
		2:08	0.50	2.50	2:32	11	2
		7:10	0.55	3.00	8:40	11	40
		2:20	0.45	1.00	0:15	12	39
		8:37	0.75	2.00	6:44	14	25
		9:00	0.50	2.50	5:23	14	33
		9:17	0.50	3.00	9:40	16	7
		9:41	0.50	3.00	8:50	16	11
		1:25	0.50	2.50	0:06	16	37
		1:44	0.50	3.00	9:56	18	13
		0:36	0.45	0.50	1:25	18	15
		9:20	0.50	3.00	9:00	18	39
		9:31	0.50	2.50	4:35	19	10
		9:41	0.50	2.50	1:05	19	40
		5:40	0.50	3.00	9:37	20	31
		9:50	0.50	3.00	8:57	21	38
		8:21	0.50	3.00	8:00	21	42
		4:58	0.50	3.00	4:30	22	15
		—	—	0.40	2:11	22	19
		8:55	0.50	3.00	5:28	23	18
		7:04	0.50	0.75	0:00	23	24
		7:52	0.50	3.00	5:50	24	31
		9:41	0.50	3.00	4:11	24	35
		9:14	0.50	3.50	5:00	25	2
		0:51	0.50	3.75	2:01	25	12
		3:35	0.50	3.00	1:34	25	16
		1:50	0.45	3.00	3:28	25	26

Best Offers Prior to Strikes—Continued

Game	Pie size	Informed		Uninformed		Player ID	
		Time	Amount	Amount	Time	Uninformed	Informed
		1:40	0.55	3.00	9:50	26	3
		8:16	0.50	3.00	9:56	26	7
		8:10	0.50	3.00	9:56	26	11
		7:54	0.50	3.00	9:52	27	8
		9:51	0.50	3.00	9:31	27	34
		9:59	0.50	3.00	6:14	29	8
		8:50	0.50	3.00	5:31	29	20
		8:50	0.50	2.00	1:52	30	9
		9:51	0.50	2.00	0:56	30	21
		9:57	0.50	2.00	6:57	31	6
		9:59	0.50	1.00	1:18	31	10
		9:52	0.50	2.00	3:21	31	18
		2:48	0.50	—	—	31	22
		7:10	0.50	4.00	6:20	31	26
		9:41	0.50	3.00	8:18	31	32
		9:12	0.50	3.00	8:52	32	23
		5:03	0.50	3.50	8:14	33	12
		2:50	0.45	3.00	2:21	33	20
		5:49	0.50	2.00	7:30	33	34
		9:50	0.50	2.50	7:15	33	38
		1:24	0.50	3.00	6:38	33	42
		0:46	0.55	0.00	0:07	34	3
		5:21	0.50	3.00	6:19	34	13
		4:21	0.50	2.00	3:17	34	19
		7:52	0.50	3.00	6:58	34	25
		9:07	0.50	3.00	8:19	34	31
		2:12	0.50	2.50	5:33	35	30
		9:47	0.50	3.00	8:51	36	1
		9:50	0.50	3.00	9:32	39	20
		7:14	0.50	2.00	4:20	39	34
		0:40	0.50	3.00	9:27	41	2
		8:52	0.50	3.00	8:30	41	18
II	big (\$3.90)	0:02	1.30	—	—	1	2
		9:59	1.20	1.95	0:14	7	4
		5:03	1.38	1.80	1:00	9	2
		0:09	1.95	2.00	8:51	12	17
		1:11	1.50	1.95	9:15	16	19
		1:35	1.15	1.90	9:42	18	9
	small (\$2.30)	3:04	1.15	1.85	2:23	4	1
		9:37	1.15	1.75	0:57	6	17
		2:26	1.15	1.85	3:36	8	5
		8:51	1.15	1.75	0:29	8	19
		9:55	1.15	1.95	8:53	17	22
III	big (\$4.20)	6:20	1.40	2.10	9:00	3	12
		0:11	2.00	—	—	7	8
		9:41	1.40	1.60	0:20	12	19
		2:00	1.37	2.10	8:48	19	10
	small (\$2.80)	1:04	1.40	2.10	8:00	1	10
		7:45	1.40	2.10	9:00	5	24
		0:52	1.70	2.10	8:25	13	22
IV	big (\$6.80)	0:07	1.50	1.80	1:37	20	17
	small (\$2.40)	4:13	1.20	1.90	0:53	1	6
		2:43	1.20	1.50	0:18	10	7

Best Offers Prior to Strikes—Continued

Game	Pie size	Informed		Uninformed		Player ID	
		Time	Amount	Amount	Time	Uninformed	Informed
		9:55	1.20	3.00	8:38	10	13
		1:07	1.20	2.00	3:41	16	7
		1:11	1.20	1.90	:37	16	11
		9:22	1.20	3.30	4:21	16	13
		0:05	1.20	2.40	4:20	17	12
		0:06	0.85	3.00	3:47	19	16
		9:38	1.20	1.20	0:00	19	24
		0:20	1.16	3.00	8:40	24	11

Note: Times are stated in minutes and seconds remaining before the end of the bargaining period. We omitted one observation because the proposals of the uninformed player were lost (this was a strike in game 1a with $\pi = \$1.00$).

REFERENCES

- Binmore, K., Shaked, A. and Sutton, J., "Fairness or Gamesmanship in Bargaining?—An Experimental Study," London School of Economics Discussion Paper 84/102, 1984.
- _____, _____ and _____, "Testing Non-cooperative Bargaining Theory: A Preliminary Study," *American Economic Review*, December 1985, 75, 1178–80.
- Cramton, Peter C., "Bargaining with Incomplete Information: An Infinite Horizon Model with Two-Sided Uncertainty," *Review of Economic Studies*, October 1984, 51, 579–94.
- Farrell, Joseph and Gibbons, Robert, "Cheap Talk Can Matter in Bargaining," *Journal of Economic Theory*, June 1989, 48, 221–37.
- Forsythe, Robert, Kennan, John and Sopher, Barry, "Dividing a Shrinking Pie: An Experimental Study of Strikes in Bargaining Games with Complete Information," in R. Mark Isaac, ed., *Research in Experimental Economics*, Greenwich, CT: JAI Press, 1990 (forthcoming).
- Fudenberg, Drew and Tirole, Jean, "Sequential Bargaining with Incomplete Information," *Review of Economic Studies*, April 1983, 50, 221–48.
- Güth, W., Schmittberger, R. and Schwarze, B., "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization*, December 1982, 3, 367–88.
- Holmstrom, Bengt and Myerson, Roger B., "Efficient and Durable Decision Rules with Incomplete Information," *Econometrica*, November 1983, 51, 1799–1819.
- Kennan, John, "The Economics of Strikes," in Orley Ashenfelter and Richard Layard, eds., *Handbook of Labor Economics*, Amsterdam: North-Holland, 1986.
- _____, "Bargaining and Strikes with Two-Sided Private Information," unpublished manuscript, University of Iowa, August 1987.
- _____, and Wilson, Robert, "Strategic Bargaining Models and Interpretation of Strike Data," *Journal of Applied Econometrics*, December 1989, 4 (supplement), S87–S130.
- Malouf, Michael W. K. and Roth, Alvin E., "Disagreement in Bargaining," *Journal of Conflict Resolution*, June 1981, 25, 329–84.
- Myerson, Roger B., "Incentive Compatibility and the Bargaining Problem," *Econometrica*, January 1979, 47, 61–73.
- _____, "Two-Person Bargaining Problems With Incomplete Information," *Econometrica*, March 1984, 52, 461–88.
- Neelin, Janet, Sonnenschein, Hugo and Spiegel, Matthew, "A Further Test of Noncooperative Bargaining Theory: Comment," *American Economic Review*, September 1988, 78, 824–36.

- Ochs, Jack and Roth, Alvin E., "An Experimental Study of Sequential Bargaining," *American Economic Review*, June 1989, 79, 355-84.
- Roth, Alvin E., "Laboratory Experimentation in Economics: A Methodological Overview," *Economic Journal*, December 1988, 98, 974-1031.
- ____ and Malouf, Michael W. K., "Game-Theoretic Models and the Role of Information in Bargaining," *Psychological Review*, November 1979, 86, 574-94.
- ____, _____ and Murnighan, Keith J., "Sociological versus Strategic Factors in Bargaining," *Journal of Economic Behavior and Organization*, June 1979, 2, 153-77.
- ____ and Murnighan, Keith J., "The Role of Information in Bargaining: An Experimental Study," *Econometrica*, September 1982, 50, 1123-42.
- ____ and Schoumaker, Francoise, "Expectations and Reputations in Bargaining: An Experimental Study," *American Economic Review*, June 1983, 73, 362-72.
- Sobel, Joel and Takahashi, Ichiro, "A Multi-Stage Model of Bargaining," *Review of Economic Studies*, July 1983, 50, 411-26.

On the Sign of the Investment–Uncertainty Relationship

By RICARDO J. CABALLERO*

Understanding the effects of uncertainty over any decision variable has fascinated economists for a long time. Risk aversion and incomplete markets are likely to make the investment–uncertainty relationship negative (e.g., Roger Craine, 1989; Joseph Zeira, 1989). What happens in the absence of risk aversion and incomplete markets is, however, ambiguous.

Richard Hartman (1972) and Andrew B. Abel (1983, 1984, 1985) found that in the presence of (symmetric) convex costs of adjustment, mean-preserving increases in price uncertainty raise investment of a competitive firm as long as the profit function is convex in prices. On the other hand, the recent literature on irreversible investment (e.g., Robert S. Pindyck, 1988; Giuseppe Bertola, 1988) has shown that increases in uncertainty lower investment. All these results have been derived under either risk neutrality or complete markets.¹

Intuition suggests that the explanation for such a difference lies with the asymmetric nature of adjustment costs in the irreversible-investment case, as compared with the symmetry of the adjustment-cost mechanisms proposed by Abel and Hartman. Although this intuition is confirmed in this paper, asymmetric adjustment costs are

shown not to be sufficient to explain why the results differ. In fact, a more hidden but at least as important difference between these two literatures is that the former assumes perfect competition and constant returns to scale, whereas the latter assumes either imperfect competition or decreasing returns to scale (or both).²

The purpose of this paper is to highlight the role of the decreasing marginal return to capital assumption (due to either imperfect competition or decreasing returns to scale [or both]) in determining the effects of adjustment-cost asymmetries on the sign of the response of investment to changes in uncertainty (under risk neutrality). For this, the paper develops a simple model with a cost-of-adjustment mechanism general enough to consider both symmetric-convexity and irreversibility as special cases. One of the most important findings is the lack of robustness of the negative relationship between investment and uncertainty under asymmetric adjustment costs³ to changes in the degree of competition. In fact, when firms are nearly competitive, the conclusion of Hartman and of Abel holds no matter how asymmetric adjustment costs are. Studying adjustment-cost mechanisms has a central role in understanding the dynamics of investment and its business-cycle implications, but conclusive results about the sign of the instantaneous relationship between uncertainty and investment should not be

*Department of Economics, Columbia University, New York, NY 10027. I am grateful to Giuseppe Bertola, Prajit Dutta, Glen Hubbard, Anil Kashyap, Richard Lyons, and the referees for their useful comments.

¹The financial literature on investment has considered risk aversion through a premium in the discount rate determined by the CAPM, (capital asset pricing model), intertemporal CAPM, or consumption CAPM. However, often this discount rate is left unchanged when studying the response of investment to uncertainty changes (e.g., Pindyck, 1988 pp. 974–5), thereby omitting the effect of changes in uncertainty on investment due to risk aversion (and incomplete markets).

²In the typical version of the irreversible-investment problem, there is no cost of upward adjustments; thus, imperfect competition and (or) decreasing returns to scale are required to bound the size of the firm.

³In this paper, *asymmetric* adjustment cost refers to the case in which it is more expensive to adjust downward than upward. Certainly, the opposite case is a trivial extension of the case studied in this paper.

expected from the adjustment-costs literature alone.

Section I develops the minimum framework necessary to understand the main issues involved. For this, a simple partial-equilibrium two-period model in which adjustment costs are (weakly) convex and possibly asymmetric and in which managers (owners) are risk-neutral is derived. Section II specializes the previous model to perfect competition. The result of Hartman and of Abel is shown to be robust to asymmetries in the adjustment-costs function, including the irreversible investment case. Hence, investment and uncertainty are positively correlated even in the extreme case of irreversible investment, as long as the firm faces a very elastic demand curve (and returns to scale are nondecreasing). Section III confirms the fact that the combination of important degrees of imperfect competition and adjustment-costs asymmetry may reverse the positive correlation between uncertainty and investment. In obtaining this result, imperfect competition is not only necessary, but is also the paramount factor. Section IV summarizes the results and discusses the roles of increasing and decreasing returns to scale. The latter makes a negative investment-uncertainty relationship more likely, whereas increasing returns makes it less likely. Finally, the Appendix presents an infinite-horizon version of the perfect-competition-adjustment-costs model.

I. Basic Framework

Each firm is in place for two periods⁴ and faces an isoelastic demand function:

$$(1) \quad P_t = Q_t^{(1-\psi)/\psi} Z_t$$

where ψ ($\psi \geq 1$) is a markup coefficient that takes the value of 1 under perfect competition, P and Q are respectively the price and quantity of the good sold, and Z is a stochastic term described by a lognormal

random-walk process:⁵

$$Z_t = Z_{t-1} \exp \varepsilon_t$$

with ε distributed normally with mean $-\sigma^2/2$ and variance σ^2 .

Technology is described by a homogeneous Cobb-Douglas production function:

$$Q = (AL^\alpha K^{1-\alpha})^\gamma$$

with A a scale parameter, L labor, K capital, α the labor share, and γ a returns-to-scale parameter.

Under these conditions the profit function, $\Pi(K, Z)$, is equal to

$$\Pi(K_t, Z_t) = h Z_t^\eta K_t^\mu$$

where

$$h = (1 - \alpha\gamma/\psi) A^{(\gamma/\psi)/(1-\alpha\gamma/\psi)}$$

$$\times \left(\frac{\alpha\gamma}{\psi w} \right)^{(\alpha\gamma/\psi)/(1-\alpha\gamma/\psi)}$$

$$\eta \equiv \frac{1}{1 - \alpha\gamma/\psi} > 1$$

$$\mu \equiv \frac{(1 - \alpha)\gamma/\psi}{1 - \alpha\gamma/\psi} \leq 1$$

and w is the (constant) wage paid to labor.⁶

Letting $C(I)$ denote the cost of changing the stock of capital by I units and assuming (without loss of generality) neither depreci-

⁵Assuming a stationary process instead of a (log) random walk does not change the conclusions.

⁶Note that the unambiguous convexity of $\Pi(\cdot, \cdot)$ with respect to Z depends on the fact that, given Q , the relation between P and Z depends on neither technology nor preference parameters. If, for example, (1) were replaced by $Q = P^{\psi/(1-\psi)} Z$, the profit function would no longer be convex with respect to Z . However, the specification in (1) seems more appropriate since it permits analysis of perfect competition as the limit case when $\psi = 1$, without altering the variance of the fundamental source of uncertainty, Z . In fact, when $\psi = 1$, (1) corresponds to the specification of Hartman and of Abel.

⁴See the Appendix for a multiperiod version of the perfect-competition version of the model.

ation nor discounting yields the following two-period optimization problem for a single firm:

$$(2) \quad V_1(K_0, Z_1) = \max_{I_1} \Pi(K_1, Z_1) - C(I_1) \\ + E_1[V_2(K_1, Z_2)] \\ \text{subject to } K_1 = K_0 + I_1$$

where V_i represents the value function at time i .

The first-order condition of this problem is

$$(3) \quad \Pi_{K_1}(K_0 + I_1, Z_1) - C_I(I_1) \\ + E_1[V_{2K_1}(K_0 + I_1, Z_2)] = 0.$$

Finally, the second-period (terminal) value function is just

$$(4) \quad V_2(K_1, Z_2) = \max_{I_2} \Pi(K_1 + I_2, Z_2) \\ - C(I_2).$$

The remainder of this section presents a general investment-cost function, while the rest of the paper discusses the role of competition, returns to scale, and the shape of the investment-cost function in determining the investment-uncertainty relationship.

The cost of changing the stock of capital by I units, denoted by $C(I)$, includes both direct and adjustment costs:

$$C(I) = I + [I > 0]\gamma_1 I^\beta + [I < 0]\gamma_2 |I|^\beta$$

where $\beta \geq 1$, γ_1 and γ_2 are two nonnegative parameters, and the price of capital has been set equal to 1.⁷

This parameterization of $C(I)$ is quite general. For example, except for the addition of I to reflect the direct cost of capital, the symmetric adjustment-cost case used by Abel (1983) is achieved when $\gamma_1 = \gamma_2 > 0$ and $\beta > 1$, and the irreversible-investment

case of Pindyck (1988) and Bertola (1988) corresponds to the case in which $\gamma_1 = 0$, $\gamma_2 = \infty$, and $\beta = 1$.

II. Perfect Competition

In order to isolate the role of competition, I will postpone issues of returns to scale until Section IV. For now, the technology is assumed to exhibit homogeneity of degree one with respect to capital and labor ($\gamma = 1$).

Moreover, perfect competition is taken only as an expository device to illustrate the consequences of a highly elastic demand. Indeed, Pindyck (1990) provides compelling arguments against mean-preserving changes in price-uncertainty experiments when competition is strictly perfect and investment is fully irreversible.

When competition is perfect $\mu = 1$; hence, the profit function becomes linear with respect to the stock of capital. This yields a simple first-order condition at time 2 [see eq. (4)]:

$$hZ_2^\eta - [I_2 > 0](1 + \gamma_1 \beta I_2^{\beta-1}) \\ - [I_2 < 0](1 - \gamma_2 \beta |I_2|^{\beta-1}) = 0.$$

Thus, I_2 is determined by

$$I_2 = \begin{cases} \left[\frac{hZ_2^\eta - 1}{\gamma_1 \beta} \right]^{1/(\beta-1)} & \text{for } I_2 > 0 \text{ or } hZ_2^\eta \geq 1 \\ - \left[\frac{1 - hZ_2^\eta}{\gamma_2 \beta} \right]^{1/(\beta-1)} & \text{for } I_2 < 0 \text{ or } hZ_2^\eta < 1. \end{cases}$$

The most important feature of this solution is that it does not depend on K_1 ; hence, the value function at time 2 is only linearly linked to K_1 through the profit function. It is easy to show that in this case $V_{2K_1} = hZ_2^\eta$.⁸

⁸The fact that $V_{2K_1} = hZ_2^\eta$ can be easily proved by noticing that I_2 does not depend on the stock of capital at time 1; therefore, $V_{2K_1} = \Pi_{K_1}(K_2, Z_2)$. Given that $\Pi(K_2, Z_2) = hZ_2^\eta K_2$ and $K_2 = K_1 + I_2$, then $V_{2K_1} = hZ_2^\eta$.

⁷Relaxing this assumption is trivial.

Plugging the expressions for $V_{2K_1}(\cdot, \cdot)$, $\Pi_K(\cdot, \cdot)$, and $C_I(\cdot)$ into equation (3) provides the first-order condition for the perfectly competitive, constant-returns-to-scale firm (at time 1):

$$\begin{aligned} & hZ_1^\eta(1 + e^{[\eta(\eta-1)/2]\sigma^2}) \\ & - [I_1 > 0](1 + \gamma_1\beta I_1^{\beta-1}) \\ & - [I_1 < 0](1 - \gamma_2\beta |I_1|^{\beta-1}) = 0 \end{aligned}$$

which yields the following investment function at time 1:

$$I_1 = \begin{cases} \left[\frac{hZ_1^\eta(1 + e^{[\eta(\eta-1)/2]\sigma^2}) - 1}{\gamma_1\beta} \right]^{1/(\beta-1)} & \text{for } I_1 > 0 \text{ or } hZ_1^\eta(1 + e^{[\eta(\eta-1)/2]\sigma^2}) \geq 1 \\ - \left[\frac{1 - hZ_1^\eta(1 + e^{[\eta(\eta-1)/2]\sigma^2})}{\gamma_2\beta} \right]^{1/(\beta-1)} & \text{for } I_1 < 0 \text{ or } hZ_1^\eta(1 + e^{[\eta(\eta-1)/2]\sigma^2}) < 1. \end{cases}$$

Again, investment at time 1 does not depend on either past or future capital stocks. This lack of "intertemporal links" does not depend on the two-period assumption. In fact, this insight also holds for the n -period model (see Appendix) and is crucial in determining the irrelevance of the shape (besides convexity) of the investment-cost function, under partial equilibrium and risk neutrality, vis-à-vis the response of investment to changes in the level of uncertainty.

The asymmetry of adjustment costs has nothing to do with the sign of the response of investment to increases in uncertainty. Whether investment is positive or negative depends on the sign of the numerator, and this does not include the adjustment-cost parameters. An increase in uncertainty raises investment (or reduces disinvestment) for any (finite) level of adjustment costs. The asymmetry determines only that investment and disinvestment have different speeds of adjustment. This is fully consistent with Hartman's (1972) and Abel's (1983,

1984, 1985) conclusion for the symmetric case. Notice that this is true even for the case in which β is very close to 1, γ_1 is slightly greater than 0, and γ_2 is ∞ , that is, when investment is irreversible and there are almost no costs (besides the price itself) of adjusting the capital stock upward.⁹

In sum, the fact that asymmetric costs imply a larger disequilibrium (as compared to the frictionless capital stock) when demand realizations are low, does *not* affect the conclusion that, under constant returns to scale and perfect competition (as well as risk neutrality and partial equilibrium), increases in uncertainty raise investment. This is just a reflection of the fact that, under perfect competition, how much is invested today affects profits tomorrow, but not the level of investment tomorrow. Therefore, any increase in the expected marginal profitability of capital, including the one caused by an increase in price uncertainty, raises investment today.

The next section relaxes the perfect-competition assumption to show that the interaction between decreasing marginal profitability of capital, resulting from imperfect competition (or decreasing returns to scale), and asymmetric costs can generate a negative investment-uncertainty relationship. In determining this relationship, imperfect competition not only is necessary but also plays a central role.

III. Imperfect Competition

When competition is imperfect there is, in general, no closed-form solution for the investment function (even in the simple two-period context).¹⁰ The exceptions correspond to extreme assumptions about $(\gamma_1, \gamma_2, \beta)$: no adjustment costs $(0, 0, \beta)$, irreversible investment with no (adjustment) cost of increasing the stock of capital $(0, \infty, 1)$, capacity constraints with no cost of

⁹It is well known that, under perfect competition and constant returns to scale, the size of the firm is indeterminate; therefore, some convexity in (upward) adjustment costs ($\gamma_1 > 0, \beta > 1$) is required in order to bound (positive) investment.

¹⁰At least, there is none known to me.

scrapping capital $(\infty, 0, 1)$, and predetermined capital (∞, ∞, β) . Here, I solve numerically several examples of the general $(\gamma_1, \gamma_2, \beta)$ case.

To streamline the notation, it is convenient to assume (without loss of generality) that $Z_1 \equiv 1$. It is also simpler (again, without loss of generality) to assume that there is no initial capital, with the result that I_1 is always positive. With these assumptions, I_1 and I_2 are determined by the following two equations:

$$\mu h I_1^{\mu-1} - \gamma_1 \beta I_1^{\beta-1} + \beta E_1([I > 0] \gamma_1 I_2^{\beta-1} - [I < 0] \gamma_2 |I_2|^{\beta-1}) = 0$$

$$\mu h Z_2^\eta (I_1 + I_2)^{\mu-1} - \beta \{ [I > 0] \gamma_1 I_2^{\beta-1} - [I < 0] \gamma_2 |I_2|^{\beta-1} \} = 0$$

corresponding to the first-order conditions in periods 1 and 2, respectively.

The problem is even further simplified by replacing the assumption of a lognormal distribution for Z with a simple symmetric Bernoulli distribution.¹¹ Figure 1 illustrates the independent (of asymmetries) role of imperfect competition in generating the negative investment-uncertainty relationship.¹² In this figure, adjustment costs are entirely symmetric; however the Jensen's inequality argument of Hartman (1972) and Abel (1983, 1984, 1985) becomes less significant as competition becomes more imper-

fect (ψ gets larger).¹³ The figure shows that as ψ increases (i.e., as the elasticity of demand is reduced), investment responds less and less to changes in the level of uncertainty.¹⁴ In fact, when $\psi > 1.6$, the lines characterizing the investment-uncertainty trade-off in Figure 1 are almost horizontal, indicating practically no effects of changes in the level of uncertainty on investment decisions. Again, it should be remembered that this has been achieved under perfectly symmetric adjustment costs.

There are two channels for the dampening of the result of Hartman and of Abel under imperfect competition (and symmetric adjustment costs). First, as the elasticity of demand is reduced, the convexity of the marginal profitability of capital with respect to price uncertainty, η , is reduced (given the stock of capital).¹⁵ This can be seen more clearly by using perfect competition as a benchmark. Recall that under perfect competition $P_t = Z_t$; hence, an increase in Z_t raises revenues both directly (through $Q\Delta P$) and indirectly through the increase in optimal output (given the stock of capital). The latter effect is responsible for the convexity of the profit function with respect to Z_t . When the elasticity of demand is less than infinite, however, the indirect effect is less important, as the firm's desire to increase output is less than that under perfect competition since doing it brings the price of its goods down, lowering the direct effect of a positive change in Z . Thus, as the elasticity of demand falls, the profit function becomes less convex with respect to Z_t . Second, as the markup (ψ) rises, the marginal profitability of capital (Π_K) decreases more with a given increase in capital (i.e., $\Pi_{KK\psi} < 0$).¹⁶ This, again, dampens the response of investment to an increase in

¹¹The interesting symmetric Bernoulli case is that in which the positive shock leads to positive investment in period 2 and a negative shock leads to disinvestment. The numerical problem is then trivial, as it consists of three equations (the expected value equation [first equation] and the two equations for the second period: one for the good realization and one for the bad realization) and three unknowns (investment in the first period and investment in the second period for the good and bad realizations). Certainly, generalizing this to any discrete-state space is trivial.

¹²All the curves (in all diagrams) are normalized by their respective level of investment under certainty.

¹³Notice that the level of adjustment costs is not important, since investment is normalized by the level of investment under certainty.

¹⁴Remember that $\psi = 1$ corresponds to the perfect competition case, whereas $\psi > 1$ represents imperfect competition (or decreasing returns to scale).

¹⁵Remember that η is decreasing in ψ .

¹⁶Remember that under perfect competition $\Pi_{KK} = 0$.

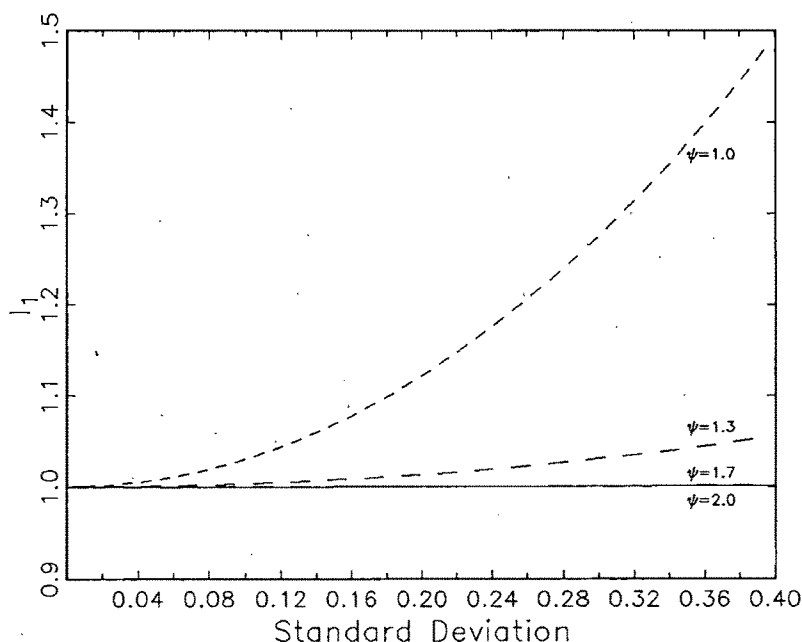


FIGURE 1. INVESTMENT AS A FUNCTION OF UNCERTAINTY FOR DIFFERENT VALUES OF ψ

price (demand) uncertainty, as the initial lift of the second period's expected marginal profitability of capital (due to Jensen's inequality) is curtailed by an increase in investment today. Combined, these two effects demonstrate that the result of Hartman and of Abel loses its strength under imperfect competition even when adjustment costs are symmetric.

Figure 2 shows the effects of asymmetric adjustment costs on the investment-uncertainty relationship, once competition is imperfect. The markup coefficient is 1.67 (this corresponds to an elasticity of demand equal to 2.5), and the cost of downward adjustments goes from being equal to the cost of upward adjustments of the capital stock (solid line), to being 50 times as expensive as the latter (short dashes). It is apparent that, given the presence of a significant degree of competition imperfection, the investment-uncertainty relationship becomes more negative as the adjustment-cost asymmetry gets larger.

Adjustment costs deform the relationship between realizations of the shock (innova-

tions), ε_2 , and the stock of capital at time 2 (when compared with the costless adjustment case). For example, if capital is predetermined (unchangeable both upward and downward), there is no link whatsoever between the realization of the shock and the stock of capital in the second period. In the irreversible-investment case, on the other hand, the capital stock and shocks are only linked for "good" realizations of the latter (i.e., for realizations in which the capital in place is less than the desired stock of capital). In general, for the asymmetric case, the stock of capital responds more to "good" than to "bad" realizations (i.e., realizations in which the capital in place is larger than the desired stock of capital). When competition is imperfect, the determination of what is a "good" and a "bad" shock is endogenous. It depends on how much is invested in the first period. The less the firm invests in the first period, the more likely it is to get a good shock (i.e., one in response to which the lowest adjustment cost is paid). Certainly, the cost of this strategy is less output today. When uncertainty is larger, "very

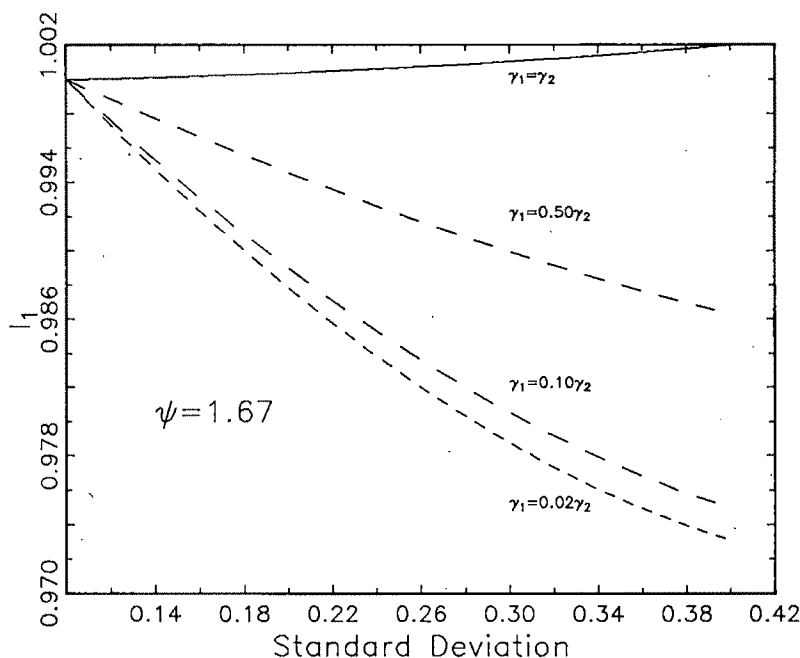


FIGURE 2. INVESTMENT AS A FUNCTION OF UNCERTAINTY FOR DIFFERENT DEGREES OF ASYMMETRY IN ADJUSTMENT COSTS ONCE COMPETITION IS IMPERFECT

good" and "very bad" news become more likely; but the larger the asymmetry of adjustment costs, the more expensive are the latter relative to the former. Thus, it is optimal to buy more protection, in the form of less initial investment, as the asymmetry and degree of uncertainty rise.

It is also worth noticing that the support of I_1 is larger in Figure 1 than in Figure 2, suggesting that imperfect competition (or decreasing returns to scale) not only is a necessary condition for asymmetric adjustment costs to affect the investment-uncertainty relationship, but also plays a central role.

IV. Conclusion

This paper has demonstrated that the presence of asymmetric adjustment costs is not sufficient to render a negative relationship between investment and mean-preserving changes in uncertainty. Some nonnegligible degree of imperfect competition is also required. In fact, the result of Hartman

(1972) and Abel (1983, 1984, 1985) (positive relationship between investment and uncertainty) for symmetric and convex adjustment costs under perfect competition fully carries over to the case of asymmetric adjustment costs. Furthermore, under very competitive conditions, the asymmetry of adjustment costs has little to do with the sign of the investment-uncertainty relationship. Today's investment decisions depend almost exclusively on the price of capital (today and in the future)¹⁷ and the expected marginal profitability of capital. In this case, the marginal profitability of capital is only tenuously related to the level of capital; hence, the convexity of marginal profitability of capital with respect to prices is the dominant factor in determining the sign of the investment-uncertainty relationship. A simple Jensen's inequality argument shows that the latter is positive.

¹⁷Assumed to be constant in this paper.

Conversely, when competition is imperfect, the marginal profitability of capital is significantly affected by the level of capital. An increase in investment today makes it more likely that the firm will find its second-period capital "too large" relative to the desired capital stock. When adjustment costs are asymmetric, having "too much" capital is worse than having "too little" of it, since increasing the stock of capital is cheaper than decreasing it. If this effect is sufficiently strong (i.e., the asymmetry of adjustment costs is large and the negative dependence of the marginal profitability of capital on the level of capital is strong), the investment-uncertainty relationship becomes negative. The irreversible-investment arguments analyzed in the literature typically correspond to this case.

Most of the analysis of this paper maintains the assumption of constant returns to scale. Relaxing this assumption, however, does not convey any additional difficulty. Convexity of the profit function with respect to prices in this case depends on the value of γ/ψ instead of just $1/\psi$. An increase in γ operates exactly like an equivalent reduction in the markup coefficient and vice versa. Hence, decreasing returns to scale makes a negative uncertainty-investment relationship more likely, whereas increasing returns offsets imperfect competition, bringing the uncertainty-investment relationship closer to the result of Hartman and of Abel.¹⁸

Overall, the results of this paper suggest that the relationship between changes in price uncertainty and capital investment under risk neutrality is not robust. Studying different adjustment-cost mechanisms is extremely important in determining the dynamics of investment and its business-cycle implications; however, it is very likely that it will be necessary to turn back to risk aver-

sion, incomplete markets, and lack of diversification to obtain a sturdier negative relationship between investment and uncertainty. Craine (1989) and Zeira (1989) have taken important steps along these lines.

APPENDIX

Consider the following discounted infinite-horizon version of the optimization problem presented in the paper for the perfect-competition case:

$$(A1) \quad V(K_{t-1}, Z_t) = \max_{I_t} \{ hZ_t^\eta K_t - I_t \\ - [I_t > 0] \gamma_1 I_t^\beta \\ - [I_t < 0] \gamma_2 |I_t|^\beta \\ + \delta E_t[V(K_t, Z_{t+1})] \}$$

subject to

$$K_t = \lambda K_{t-1} + I_t$$

$$\lim_{T \rightarrow \infty} \delta^T E_t[V(K_{T-1}, Z_T)] = 0$$

$$K_{t-1} \text{ given.}$$

The parameters δ and λ correspond to the discount factor and 1 minus the depreciation rate, respectively, both being less than 1. Also assume that $Z_t = Z_{t-1}W_t$, where W_t is any strictly positive independently and identically distributed random variable with mean 1 and log-standard deviation σ .

From the insights given by the two-periods problem, it seems reasonable to guess a value function of the form

$$(A2) \quad V(K_{t-1}, Z_t) = A(Z_t) + chZ_t^\eta K_{t-1}$$

where c is a constant and $A(\cdot)$ is a continuous function, both to be found.

After substituting (A1) into (A2) and obtaining the first-order conditions, it is possible to write investment as a function of the

¹⁸ It is also easy to show that if there are costs of waiting to invest, in the sense that there are some advantages (besides the traditional convex-adjustment-costs arguments) of planning and investing with time, the investment-uncertainty relationship may become positive even when adjustment costs are asymmetric, returns to scale are constant or decreasing, and competition is very imperfect.

unknown parameter c :

$$I_t = \begin{cases} \left[\frac{hZ_t^\eta(1 + c\lambda\delta L(\eta, \sigma)) - 1}{\gamma_1\beta} \right]^{1/(\beta-1)} & \text{for } I_t > 0 \\ - \left[\frac{1 - hZ_t^\eta(1 + c\lambda\delta L(\eta, \sigma))}{\gamma_2\beta} \right]^{1/(\beta-1)} & \text{for } I_t < 0 \end{cases}$$

where $L(\eta, \sigma) = E_t[W_{t+1}^\eta]$, an increasing function of σ (Jensen's inequality). Plugging this back into the Bellman equation makes it possible to find c :

$$c = \frac{\lambda}{1 - \lambda\delta L(\eta, \sigma)}.$$

This substitution also leads to a functional equation for $A(\cdot)$:

$$(A3) \quad A(Z_t) = \delta E_t[A(Z_{t+1})] + G(Z_t)$$

where

$$\begin{aligned} G(x) \equiv & - \left[x \geq \{h[1 + c\lambda L(\eta, \sigma)]\}^{-1/\eta} \right] \\ & \times \left[\frac{hx^\eta[1 + c\lambda\delta L(\eta, \sigma)] - 1}{\gamma_1\beta} \right]^{1/(\beta-1)} \\ & \times [1 - c\delta L(\eta, \sigma)hx^\eta] \\ & - \left[x \geq \{h[1 + c\lambda L(\eta, \sigma)]\}^{-1/\eta} \right] \\ & \times \gamma_1 \left[\frac{hx^\eta[1 + c\lambda\delta L(\eta, \sigma)] - 1}{\gamma_1\beta} \right]^{\beta/(\beta-1)} \\ & + \left[x < \{h[1 + c\lambda L(\eta, \sigma)]\}^{-1/\eta} \right] \\ & \times \left[\frac{1 - hx^\eta[1 + c\lambda\delta L(\eta, \sigma)]}{\gamma_2\beta} \right]^{1/(\beta-1)} \\ & \times [1 - c\delta L(\eta, \sigma)hx^\eta] \\ & - \left[x < \{h[1 + c\lambda L(\eta, \sigma)]\}^{-1/\eta} \right] \\ & \times \gamma_2 \left[\frac{1 - hx^\eta[1 + c\lambda\delta L(\eta, \sigma)]}{\gamma_2\beta} \right]^{\beta/(\beta-1)}. \end{aligned}$$

Making the simplifying assumption that W_t has bounded support makes the function $G(x)$ a continuous bounded function. This, together with the Markov structure of the transition and the fact that $\delta < 1$, determines that $A(Z_t)$ exists and is unique (David Blackwell, 1965).¹⁹

As long as $\lambda\delta L(\eta, \sigma) < 1$, the instantaneous reward function is concave in the control I_t ; therefore, the investment function is also unique. Furthermore,

$$\begin{aligned} \frac{\partial I_t}{\partial \sigma} &= \{\beta(\beta-1)([I > 0]\gamma_1 + [I < 0]\gamma_2) \\ &\quad \times (1 - \lambda\delta L(\eta, \sigma))^2\}^{-1} \\ &\quad \times |I_t|^{-1/\beta} hZ_t^\eta \delta \lambda^2 L_\sigma(\eta, 0) \\ &> 0 \end{aligned}$$

confirming the fact that the result of Hartman (1972) and Abel (1983, 1984, 1985) extends to the asymmetric-adjustment-cost case.

¹⁹Note that the function $A(\cdot)$ does not need to be unique to guarantee a unique investment function.

REFERENCES

- Abel, Andrew B., "Optimal Investment Under Uncertainty," *American Economic Review*, March 1983, 73, 228-33.
- , "The Effects of Uncertainty on Investment and the Expected Long-Run Capital Stock," *Journal of Economic Dynamics and Control*, February 1984, 7, 39-54.
- , "A Stochastic Model of Investment, Marginal Q and the Market Value of the Firm," *International Economic Review*, June 1985, 26, 305-22.
- Bertola, Giuseppe, "Adjustment Costs and Dynamic Factor Demands: Investment and Employment Under Uncertainty," Ph.D. Dissertation (Ch. 2), MIT, June 1988.
- Blackwell, David, "Discounted Dynamic Programming," *Annals of Mathematical*

Statistics, February 1965, 36, 226-35.

Craine, Roger, "Risky Business: The Allocation of Capital," *Journal of Monetary Economics*, March 1989, 23, 201-18.

Hartman, Richard, "The Effects of Price and Cost Uncertainty on Investment," *Journal of Economic Theory*, October 1972, 5, 258-66.

Pindyck, Robert S., "Irreversible Investment,

Capacity Choice, and the Value of the Firm," *American Economic Review*, December 1988, 78, 969-85.

———, "A Note on Competitive Investment Under Uncertainty," mimeo, MIT, June 1990.

Zeira, Joseph, "Cost Uncertainty and the Rate of Investment," Hebrew University Working Paper No. 206, February 1989.

Savings and Wealth in Models with Altruistic Bequests

By WILLIAM LORD AND PETER RANGAZAS*

During recent years much attention has been given to the role of bequests in explaining various aspects of economic behavior. For example, authors have studied the impact of bequests on physical wealth accumulation (Laurence Kotlikoff and Lawrence Summers, 1981), the interest elasticity of savings (Owen Evans, 1983), tax reform (Laurence Seidman, 1984), and the distribution of wealth (Alan Blinder, 1976a).

More recently, economists have begun to distinguish between human and physical bequests (Gary Becker and Nigel Tómes, 1986). Acknowledging the presence of bequest in human form is important in analyzing each of the issues mentioned above. Outlays on children affect the shape of the family's labor supply and expenditure profiles, upon which the level of aggregate savings crucially depends. If these outlays are interest-sensitive substitutes for physical bequests, they will also influence the interest elasticity of savings. In the area of tax reform, the neutrality of the consumption tax depends on the ability of the government to identify human capital bequests from general consumption (Raymond Batina, 1987). If this is not possible, the "portfolio-choice" of investment in human versus physical capital may be distorted at the margin. Finally, the unequal transmission of wealth across generations may be explained by the lack of physical bequests and the inefficiently low levels of human bequests made by wealth-constrained households (Becker and Tómes, 1986). To distinguish clearly the behavior of constrained households from unconstrained households, who make sizeable physical be-

quests, both types of bequests must be recognized.¹

This paper introduces a model that makes explicit examination of these issues possible. It extends Lord's (1989) multiperiod model of life-cycle savings and adult human capital investment by adding altruistically motivated human and physical bequests.² The model is calibrated using microeconomic data on earnings, time and goods expenditures on children's human capital, and physical bequests. We then use the model to examine the contribution of bequests to wealth accumulation and the level of savings. This topic is the source of an important but yet unresolved debate between Franco Modigliani (1988) and Kotlikoff (1988). Kotlikoff maintains that U.S. wealth accumulation is primarily the consequence of bequests, as opposed to life-cycle savings for retirement. One type of evidence cited by Kotlikoff comes from simulation results demonstrating the failure of realistically calibrated life-cycle models to generate sufficiently high saving rates and wealth:income ratios (Alan Auerbach and Kotlikoff, 1987). We show, however, that if a pure life-cycle

¹In addition to these theoretical considerations, human bequests are worthy of study because they are likely to be much larger than financial bequests in the aggregate (Blinder, 1976b p. 90). We form estimates of the relative shares of wealth devoted to human and physical bequests in the Appendix.

²Evans (1983) and Seidman (1983, 1984) augment the standard life-cycle savings model with physical, but not human, bequests. Evans briefly considers a simple dynastic model of altruistic bequests but concentrates primarily on ad hoc bequest formulations which are not derived from utility maximization. Seidman uses the "taste-for-bequest" model, in which the source of satisfaction is the level of bequest per se. This formulation immediately encounters difficulties when both human and physical bequests are included. Unless the price of the inputs in human capital formation equals the price of physical bequests, the household will be at a corner, leaving all physical or all human bequests, depending on which is cheaper.

*Department of Economics, University of Maryland-Baltimore County, Catonsville, MD 21228 and Department of Economics, IUPUI, Indianapolis, IN 46202-5140. We are grateful to Sharon Zehr Rangazas for her assistance and comments and to the referees for many useful suggestions.

model generates low aggregate savings rates, then augmenting the model with altruistic bequests that mimic available microdata will not necessarily cause the savings rate to rise significantly.

I. The Model

The general altruistic approach to bequests has been developed most recently by Becker (1981) and Becker and Tomes (1986).³ It presents a natural and parsimonious theory of the case in which both human and physical bequests are made. Our model is a special case of this paradigm designed to facilitate comparisons to the standard life-cycle simulation model found in the literature (Summers, 1981; Evans, 1983; Auerbach and Kotlikoff, 1987).

We begin by describing the age structure, wealth constraint, and production technologies of the model. Individuals are economically dependent on their parents through age 20. At age 21, they begin their independent economic lives. At age 26, the unisex individual "produces" 1.3 children, which corresponds to a population growth rate of 1 percent. Retirement occurs at age 63, and individuals die after 78 years of life. At this time, physical bequests are left to the next generation. Inheritance is, therefore, received at age 53.

Each individual begins his economic life with an initial quantity of nondepreciating human capital,⁴ based on his parent's investment, which augments his adult human capital stock in an additive fashion. Throughout economic life, individuals choose levels of family consumption, goods and human inputs for their children's development, and the level of a physical bequest inclusive of interest received and taxes paid, as well as time and goods inputs for their own adult human capital development.

The modeling of adult human capital decisions is based on Yoram Ben-Porath (1967). During adult years, the production function for gross additions to the stock of human capital follows a Cobb-Douglas specification with decreasing returns to scale and constant output elasticities. Human capital produced during adulthood is assumed to depreciate at a constant rate, d . The technology constraining a child's human capital production is distinct from that of adults, although of the same basic Cobb-Douglas form. The human capital production function for children is

$$(1a) \quad Q_t = \nu (s_t H_t)^{\gamma_t} (D_t)^{\lambda_t}$$

$$(1b) \quad \gamma_t = \frac{\gamma_0}{(1 + \alpha)^{t - t_0}}$$

$$(1c) \quad \lambda_t = \frac{\lambda_0}{(1 + \beta)^{t - t_0}}$$

where, ν , γ_0 , λ_0 , α , and β are constants and where $t_0 = 26$, the first year of investment in children. Equation (1) allows the output elasticities of parental time and goods inputs to vary with the age of the child. This is necessary to capture certain stylized facts which are discussed in the Appendix.

For comparative purposes, we employ the familiar constant-elasticity-of-substitution utility function (Summers, 1981; Evans, 1983; Auerbach and Kotlikoff, 1987) augmented with altruistic preferences toward the next generation;

$$(2) \quad U = \frac{1}{1 - 1/\sigma}$$

$$\times \left[\sum_{t=1}^{58} C_{20+t}^{1-1/\sigma} (1 + \delta)^{-(t-1)} + m(1.01)^{26} U_* (1 + \delta)^{-57} \right]$$

where σ is the intertemporal elasticity of substitution in consumption, δ is the pure

³Paul Menchik and Martin David (1983) also discuss this model and trace its origin to Alfred Marshall.

⁴For the typical household, basic reading, writing, and arithmetic skills, as well as health habits, are likely to be maintained by the daily experiences of production and consumption. For this reason, the initial stock of human capital is not subject to depreciation.

rate of time preference, U_* is the utility of a member of the next generation, and m is the relative preference for U_* versus own lifetime consumption. The microeconomic solutions are obtained in the manner of Batina (1987). In the steady state, where $U = U_*$, the first-order conditions of this otherwise complex problem collapse to a rather simple form, including

$$(3) \quad m(1 + \delta)^{-57} = (1 + r)^{-26}.$$

A special case of this model takes $m = 1$ and assumes nonoverlapping generations, which reduces (3) to the familiar condition, $r = \delta$.

Equation (3) indicates that the net interest rate is completely determined by taste and demographic parameters, implying an infinitely elastic partial-equilibrium supply of capital. As a result, bequests, and thus savings, are indeterminate at the micro-level. To obtain a solution, values for all parameters, as well as initial values for r and bequest, are assumed. First, the efficient human capital decisions are computed. Next, the consumption choices are solved for in terms of the bequest. Consumption is then aggregated across cohorts using single-period resource constraints, to obtain an expression for the aggregate stock of assets. This, along with the aggregate effective labor supply, determines the capital:labor ratio. Substituting the capital:labor ratio into the profit-maximizing condition generates a gross interest rate. Bequests are adjusted until the interest rate generated is in agreement with the initially specified r . Successive trials lead to a choice of r best satisfying the stylized facts (see Appendix).

II. Initial Baselines and Savings Rates

Setting the parameter values of a simulation model is a critical step in the modeling process. The exact approach taken depends on the availability of empirical information and the purpose of the study. We use three different methods of setting the parameter values of our model. First, empirical estimates of the parameters are used, when

they are available and precise. The empirical literature typically produces a range of estimates for a parameter. Rather than utilize the full range, we choose to focus on the upper and lower bounds. We collect the upper-bound estimates into what is called the "high-response" (H) case, since these estimates suggest relatively large behavioral responses and high interest elasticities. The lower-bound estimates are collected in the "low-response" (L) case to reflect the other extreme.

Second, other parameters are set in order to produce stylized facts regarding descriptive aspects of household behavior, for example, the expenditure shares on various goods. Finally, to facilitate comparisons to existing models, any remaining free parameters are to be set equal to values chosen by previous authors.⁵ A detailed account of this procedure is presented in the Appendix.

We consider six baselines in total, with a high (H) and low (L) response case for each of three assumptions made about the share of lifetime wealth devoted to bequests. Our review of the available evidence yields three different estimates of the bequest share: 1, 2.5, and 4 percent (see Appendix). Table 1 presents the simulations for a number of variables under each case.

The range of the Modigliani wealth share (row 5), aggregate inheritance divided by aggregate wealth, includes the estimates obtained by Modigliani and by Kotlikoff and Summers (see Modigliani, 1988 p. 28). Using methods entirely independent from ours, Modigliani produces an estimate of 17 percent, compared to Kotlikoff and Summers's estimate of 46 percent. Thus, their estimates are consistent with the micro-bequest

⁵Since the purpose of our paper is to provide a counterexample to commonly held beliefs about the role of bequests in aggregate savings, this procedure is appropriate. This is not to say that we conduct no sensitivity analysis. Recall that we provide high- and low-response cases and, in addition, experiment with functional forms and unknown parameters as mentioned in the Appendix. However, if the study were designed to test alternative theories or to provide recommendations, a more thorough sensitivity analysis of the free parameters should be conducted.

TABLE 1—STEADY-STATE BASELINES

Case variable	Bequest share					
	1%		2.5%		4%	
	H	L	H	L	H	L
Savings rate ^a	4.0	3.5	4.2	3.8	4.3	4.0
Interest rate ^b	5.0	5.7	4.8	5.3	4.6	4.9
Marginal propensity to bequeath ^c	0.27	0.25	0.29	0.27	0.30	0.28
Expenditure share ^d	5.2	3.9	5.4	4.3	5.6	4.6
Modigliani wealth share ^e	0.14	0.27	0.32	0.49	0.47	0.62
Wealth share including interest ^f	0.26	0.54	0.59	0.96	0.83	1.15
Flow share ^g	0.01	0.01	0.01	0.02	0.02	0.03
Savings rate ^h (no human bequest)	4.2	3.6	4.3	3.9	4.5	4.2
Savings rate ⁱ (no bequests)	4.0	3.3	4.0	3.2	3.9	3.1

Note: H and L are the high- and low-response cases.

^aAggregate savings divided by aggregate output.

^bThe net of tax interest rate, with a tax rate of 20 percent.

^cIncrease in the present value of financial bequests following a one-unit increase in parents' wealth.

^dThe present value cost of time and goods inputs allocated to children's human capital divided by human wealth.

^eTotal inheritance wealth divided by total wealth.

^fAdds interest income from inheritance to Modigliani wealth share.

^gThe flow of bequests divided by total wealth.

^hAggregate savings divided by aggregate output when the efficient level of young human capital is costlessly provided.

ⁱAggregate savings divided by aggregate output when there are no human or financial bequests.

share estimates we constructed from previous cross-sectional empirical studies and simulation models. The next row (row 6) includes the potential interest income earned on inheritances in the numerator and again produces numbers comparable to the Modigliani and Kotlikoff and Summers range of estimates (see Modigliani, 1988 p. 28).

A most surprising feature of our baselines is that the aggregate savings rates (row 1) show little variation across bequest-share assumptions. The upper-bound bequest-share case, despite producing Modigliani wealth-share estimates slightly above those obtained by Kotlikoff and Summers, continues to produce a savings rate in the neighborhood of 4 percent. Furthermore, this is only marginally higher than the savings rate obtained with the bequest motive removed and all other assumptions maintained (last row). This finding contradicts the commonly held belief that life-cycle models augmented with a bequest motive would produce sig-

nificantly higher savings rates. For example, based on his work with Auerbach, Kotlikoff points out that realistically calibrated life-cycle simulation models have difficulty generating realistic savings rates, especially when the consumption of children is included (Kotlikoff, 1988 p. 48; Auerbach and Kotlikoff, 1987 pp. 64, 168). In the version of their model which excludes children, Auerbach and Kotlikoff report a savings rate of 3.7 percent, a value within the range of our no-bequests steady-state figures (last row of Table 1), where we assume that the previously endogenous level of young human capital is costlessly provided and there is no bequest motive. This rate of savings is substantially below either private or national savings of the postwar period, with the exception of the national rate in the last ten years.⁶ When they add the consumption

⁶See Michael Boskin (1988 table 3).

of children, they are also forced to add a positive government capital stock to maintain reasonable levels of aggregate wealth. They claim that this "...provides further evidence of the inadequacy of the pure life-cycle model without bequests to explain observed rates of capital accumulation" (p. 168).

Our full-model savings rates (first row), suggest that the bequest motive does offset the reduction in savings resulting from expenditures on children, without the need of government savings. However, this is misleading. As mentioned in the Appendix, we deliberately underestimate expenditures on children in order to reflect something closer to pure investment expenditures. More importantly, the model with financial bequests but no human capital expenditures on children (second to last row of Table 1), again yields only marginally higher rates of savings than the full model. Thus, in our model, the impact of financial and human bequests has very little effect on physical capital accumulation. There appears to be a substitution between the different types of savings and the different types of expenditures, so as to leave total savings rates essentially unchanged.

This can best be explained by starting with the full model and then carrying out the Kotlikoff and Summers (1981) thought experiment of eliminating the bequest motive. First, consider the elimination of human bequests, while adjusting the altruism parameter to keep the physical inheritance share constant. This has a positive wealth effect upon parents, through reduced expenditures. Consequently, increased consumption significantly dampens any increase in life-cycle savings. For the most part, consumption, (and not life-cycle savings) is substituted for human capital expenditures on children.

Second, suppose financial bequests are also removed. Again, savings rates fall only slightly. This is due to the wealth effects associated with the elimination of financial bequests. As is well known, so long as the interest rate r exceeds the (effective) population growth rate g , bequests generate positive net-of-bequest inheritance wealth. In the present case, r exceeds g by almost 4

percent, on average. Consequently, a large part of the inheritance is used to finance life-cycle consumption. As a result, eliminating financial bequests causes an increase in life-cycle savings to preserve retirement consumption, leaving savings rates virtually unchanged.

The savings rates do fall by relatively greater amounts in the low-response case. In this case, as bequests are removed and interest rates rise, the decrease in human capital expenditures and consumption is smaller than in the high-response case. As a result, the offsetting increase in life-cycle savings is smaller. Thus, substitution effects (and not wealth effects alone), play a role in the results. However, even when assuming lower-bound responses, bequests make a small contribution to aggregate savings.

III. Conclusion

It has been argued in the literature that incorporating bequests into life-cycle simulation models may raise savings rates to more reasonable levels. We have shown, for the initial levels of savings generated by realistically calibrated life-cycle models, that adding bequests will not significantly alter savings rates. Adding or subtracting human and financial bequests in a life-cycle model causes indirect wealth effects on lifetime consumption, which serve to offset the direct impact on savings. The offsetting wealth effect is relatively large when the difference between the steady-state interest rate and the effective population growth rate is relatively large. For a given population growth rate, the size of this gap is driven by the underlying level of life-cycle savings. The lower the level of life-cycle savings, the larger is the gap and the smaller is the effect on savings from altering the level of bequests.

APPENDIX

As in Lord (1989) and Auerbach et al. (1988), adult human capital production parameters are chosen to approximate an age-wage profile reported by Finis Welch (1979) and to reflect estimates of the returns to scale (James Heckman, 1976), of

human capital decay (Heckman, 1976; B. F. Kiker and Blaine Roberts, 1984), and of the time and goods input shares (Boskin, 1975). These considerations suggest two adult human capital parameter-value combinations, one with high returns and low decay (0.75 and 0.08) used in the H case and another with low returns and high decay (0.45 and 0.12) used in the L case. Both combinations have a time input share of 0.67.

A survey of the empirical literature studying family expenditures on children reveals five general stylized facts. These deal with the total lifetime budget shares to human and physical bequests, the portion of total human capital investments made in the form of goods versus time, and the profiles of the parents' time and goods inputs during the child's period of dependence.

Boone Turchi (1975) estimated the present value of parental outlays on children through the college years to be approximately 6.5 percent of human wealth. Thomas Espenshade (1984) expanded the expenditure categories used by Turchi to include approximately 50 percent more before-college goods inputs. Accounting for the expanded expenditure categories increases the total estimate to 7.6 percent of human wealth. The latter estimate also makes the shares of time and goods inputs roughly equal. Without a doubt, many of these expenditures are for consumption rather than investment purposes. Consequently, when calibrating the model, we kept the total expenditure share out of human wealth between 3 and 6 percent.

The expenditure share (in present value) of human wealth devoted to bequests can be expressed as $[(1+g)/(1+r)]^N [\pi/(1-\pi)]$, where π is the present value of inheritance as a fraction of total wealth. Let $N = T_y + 1 + n$, with T_y being the age in the last year of economic dependence and n being the number of years waited, after economic independence, before having children. Thus, N is the age gap between generations. The Survey Research Center's 1960 "Income and Welfare" survey of inheritances yields an average π , for the population, of 2.5 percent (see James Davies, 1982 pp. 473-4). Independently, Blinder (1973) and Davies

(1982), using empirically grounded micro-simulations, generate an average π of 5 or 6 percent. With g set equal to 1 percent, the after-tax interest rate set equal to 3 percent, and a value of N equal to 26 years ($T_y = 20$ and $n = 5$), these inheritance figures give a range of estimates for the bequest share between 1.5 and 3.2 percent.

Menchik and David (1983) find a bequest share of 10 percent for the top quintile in their sample. Lower quintiles in their sample also exhibit positive shares, but they are unstable in value across age cohorts and are unresponsive to changes in wealth, unlike the shares in the upper quintile. It seems likely that the majority of bequests left by the lower quintiles were accidental rather than altruistically planned. Setting the bequest shares of the lower four quintiles to zero and assuming that 40 percent of human wealth is owned by the top quintile gives an average bequest share for the population around 4 percent. In our simulations, we chose to reflect the range of available evidence by producing aggregate savings rates under three different shares: 1, 2.5, and 4 percent.

Finally, a recent study by Joseph Hotz and Robert Miller (1988) presents some information on the time profile of parents' time and goods inputs. They estimate that parents devote about 660 hours to child care in the initial year of dependency. With 16 hours of discretionary time per day, this translates to approximately 5 percent of the yearly time endowment. After the first year, they estimate the time input to decline geometrically at a rate of 12 percent. They also could not reject the hypothesis that goods inputs had a constant profile. This seems to be roughly consistent with the slightly humpbacked expenditure profile we computed from estimates in Espenshade (1984).

These facts regarding expenditures on children can be mimicked by appropriately setting the parameters of the children's human capital production function. The values for α and β [see equations (1b) and (1c)] were chosen to approximate the time and goods input profiles and relative expenditure shares. The efficiency scalars in adult and young human capital production and

the returns to scale in the children's production function were then set to generate the values of total expenditures relative to human wealth and the magnitude of adult production relative to the bequeathed stock.

The three taste parameters are the rate of time preference δ , the intertemporal substitution elasticity σ , and the taste for bequest parameter m . Recent estimates by Robert Hall (1988), using aggregate data, suggest that σ is not likely to be much above 0.10. The most recent estimates using micro-data are those obtained by Thomas MaCurdy (1981). They place an upper bound of σ at 0.45. We construct baselines, with $\sigma = 0.15$ (L case) and $\sigma = 0.4$ (H case). In our reported results, δ is set at 0.01 in all cases, a value intermediate to those assumed by Summers (1981), Evans (1983), and Auerbach and Kotlikoff (1987). We also conducted the experiments with $\delta = 0.03$, the highest value found in the literature, and none of the qualitative conclusions were altered. Evans is the only researcher to set δ below 0.01. Since most economists regard δ to be positive, we did not consider negative values. As mentioned, bequest parameters are set to produce bequests of 1, 2.5, and 4 percent of human wealth.

In addition to the altruistic model reported in the text, we considered the case in which the next generation's wealth, as opposed to utility, entered the current generation's utility function. We looked at this model with both homothetic and nonhomothetic utility functions. For the parameter settings just described, these models produced very similar results and are apparently close substitutes to the formulation used in the text for the issues at hand. In future work, we plan a more thorough examination of different altruistic specifications.

REFERENCES

- Auerbach, Alan J. and Kotlikoff, Laurence J., *Dynamic Fiscal Policy*, Cambridge, U.K.: Cambridge University Press, 1987.
- _____, Kotlikoff, Laurence J. and Skinner, Jonathan, "The Efficiency Gains from Dynamic Tax Reform," *International Economic Review*, February 1983, 24, 81-99.
- Batina, Raymond G., "The Consumption Tax in the Presence of Altruistic Cash and Human Capital Bequests with Endogenous Fertility Decisions," *Journal of Public Economics*, December 1987, 34, 329-54.
- Becker, Gary, *A Treatise on the Family*, Cambridge, MA: Harvard University Press, 1981.
- _____, and Tomes, Nigel, "Human Capital and the Rise and Fall of Families," *Journal of Labor Economics*, July 1986, 4, S1-S39.
- Ben-Porath, Yoram, "The Production of Human Capital and the Life Cycle of Earnings," *Journal of Political Economy*, August 1967, 75, 352-65.
- Blinder, Alan S., "A Model of Inherited Wealth," *Quarterly Journal of Economics*, November 1973, 87, 608-26.
- _____, (1976a) *Toward An Economic Theory of Income Distribution*, Cambridge, MA: MIT Press, 1976.
- _____, (1976b) "Intergenerational Transfers and Life Cycle Consumption," *American Economic Review*, May 1976 (*Papers and Proceedings*), 66, 87-93.
- Boskin, Michael J., "Notes on the Tax Treatment of Human Capital" in *Conference on Tax Research*, Stanford, CA, Cambridge, MA: NBER, 1975.
- _____, "Issues in the Measurement and Interpretation of Savings and Wealth," NBER (Cambridge, MA) Working Paper No. 2633, 1988.
- Davies, James B., "The Relative Impact of Inheritance and Other Factors on Economic Inequality," *Quarterly Journal of Economics*, August 1982, 47, 471-98.
- Espenshade, Thomas J., *Investing in Children*, Washington, DC: Urban Institute Press, 1984.
- Evans, Owen J., "Tax Policy, the Interest Elasticity of Savings, and Capital Accumulation: Numerical Analysis of Theoretical Models," *American Economic Review*, June 1983, 73, 398-409.
- Hall, Robert E., "Intertemporal Substitution in Consumption," *Journal of Political Economy*, April 1988, 96, 339-57.

- Heckman, James, "A Life-Cycle Model of Earnings, Learning and Consumption," *Journal of Political Economy*, August 1976, 84, S11-S44.
- Hotz, Joseph V. and Miller, Robert A., "An Empirical Analysis of Life Cycle Fertility and Female Labor Supply," *Econometrica*, January 1988, 56, 91-118.
- Kiker, B. F. and Roberts, R. Blaine, "The Durability of Human Capital: Some New Evidence," *Economic Inquiry*, April 1984, 22, 269-81.
- Kotlikoff, Laurence J., "Intergenerational Transfers and Savings," *Journal of Economic Perspectives*, Spring 1988, 2, 41-58.
- _____, and Summers, Lawrence H., "The Role of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, 2, 706-32.
- Lord, William, "The Transition From Payroll to Consumption Receipts with Endogenous Human Capital," *Journal of Public Economics*, February 1989, 38, 53-74.
- MaCurdy, Thomas E., "An Empirical Model of Labor Supply in a Life Cycle Setting," *Journal of Political Economy*, December 1981, 89, 1059-85.
- Menchik, Paul L. and David, Martin, "Income Distribution, Lifetime Savings, and Bequests," *American Economic Review*, September 1983, 83, 672-90.
- Modigliani, Franco, "The Role of Intergenerational Transfers and Life Cycle Savings in the Accumulation of Wealth," *Journal of Economic Perspectives*, Spring 1988, 2, 41-58.
- Seidman, Laurence, "Taxes in a Life Cycle Growth Model with Bequests and Inheritances," *American Economic Review*, June 1983, 73, 437-91.
- _____, "Conversion to a Consumption Tax: The Transition in a Life-Cycle Growth Model," *Journal of Political Economy*, June 1984, 92, 247-67.
- Summers, Lawrence, "Capital Taxation and Accumulation in a Life-Cycle Growth Model," *American Economic Review*, September 1981, 71, 533-44.
- Turchi, Boone A., *The Demand for Children: The Economics of Fertility in the United States*. Cambridge, MA: Ballinger, 1975.
- Welch, Finis, "Effects of Cohort Size on Earnings: The Baby Boom's Financial Bust," *Journal of Political Economy*, October 1979, 87, S65-S97.

A New Estimate of the Welfare Loss of Excess Health Insurance

By ROGER FELDMAN AND BRYAN DOWD*

In a classic article, Martin S. Feldstein (1973) argued that American families are in general overinsured against medical care expenses. If insurance were reduced, the utility loss from increased risk would be more than outweighed by the gain from lower prices and reduced purchase of excess medical care. Raising the average coinsurance rate from 0.33 to 0.67 would produce a net gain of \$27.8 billion (in 1984 prices) in the private hospital sector, according to Feldstein's "most likely" parameter values.

To make his estimate, Feldstein assumed values for the following parameters: the price elasticity of demand, consumer risk-aversion, the gross price change resulting from lower insurance coverage, and the decrease in health-care quality induced by lower insurance coverage. While the actual values of the last two parameters are still somewhat problematic, more recent and much-improved estimates of the price elasticity of health-care demand and risk-aversion are now available. The Rand Health Insurance Experiment has shown that consumers randomly assigned to a free health-care plan spent 46 percent more than those in a plan with 95 percent coinsurance (Willard G. Manning et al., 1987). The price elasticity of demand for a constant coinsurance policy is in the -0.1 to -0.2 range, values that are at the lower end of those in the nonexperimental literature, and less elastic than the -0.4 to -0.8 range used by Feldstein.

Improved estimates of the degree of consumer risk-aversion were also produced

by the Health Insurance Experiment (M. Susan Marquis and Martin R. Holmer, 1986). Families in all plans (except the free-care plan) faced a maximum limit on out-of-pocket expenditures in any year. At the end of its participation in the experiment, each family was asked if it would be willing to pay a predetermined premium to reduce this limit. From the resulting offers, one can calculate the Pratt-Arrow absolute risk-aversion parameter (John W. Pratt, 1964; Kenneth J. Arrow, 1974). Similar estimates are available from Bernard Friedman's (1974) study of health-plan choice by Federal employees. These data suggest that consumers are considerably more risk-averse than Feldstein assumed.

The purpose of this paper is to calculate new estimates of the welfare loss of excess health insurance, using the Rand data on price elasticity and consumer risk-aversion. Manning et al. (1987) reported the dead-weight loss from insurance in a recent issue of this journal. According to Manning et al. (1987), individuals in the free-care plan spent \$344 more, on average, than those in the 95-percent coinsurance plan. The welfare gain per person from reducing insurance coverage was calculated at approximately $\$344(1 - 0.31/2) = \291 (in 1984 prices), where 0.31 is the average coinsurance rate.¹

These calculations assume that gross medical-care prices are constant. This assumption affects not only Manning et al.'s (1987) estimate of the welfare loss of excess health insurance, but also their explanation

*University of Minnesota, School of Public Health, Division of Health Services Research and Policy, 420 Delaware Street SE, Box 729, Minneapolis, MN 55455. Feldman is also in the Department of Economics, University of Minnesota. The authors acknowledge helpful comments from Will Manning.

¹Because the maximum limit on out-of-pocket expenses in the 95-percent plan was capped at \$1,000, these individuals faced an average coinsurance rate of 0.31. An upper bound on the welfare-loss calculation comes from assuming that they valued the last dollar of medical-care spending at 31 cents.

of the relationship between increased health insurance and health-care expenditures over the period from 1950 to 1980. It is unlikely that gross price would remain constant if health insurance coverage were altered on a national scale. Feldstein (1971, 1973) suggests that gross price would fall if coverage were reduced, although the exact size of the price decrease is uncertain.

Further, a change in health insurance might induce changes in product quality. In Feldstein's model, the hospital changes quality by changing the input mix. Manning et al. (1987) cite a different effect of insurance: its stimulus on technological change. However, the extent to which insurance induces changes in input mix and technology is not clear, nor is the effect of these changes on consumer demand. Consequently, we simulate several changes in gross medical-care prices in our analysis, including the baseline case of constant prices.

Finally, in concentrating on the dead-weight loss from moral hazard, Manning et al. (1987) did not consider the benefits of insurance, namely, the reduction in financial risk. While not taking issue with Manning et al.'s (1987) basic conclusions—that health insurance encourages excess consumption, but that health insurance cannot account for much of the increase in total health care spending at constant prices—we suggest that the welfare-loss calculation can be refined to include gross price and risk-bearing effects.

The exact welfare-loss calculation depends explicitly on the utility function and the statistical distributions of expenditures for outpatient visits, hospital admissions, and length of stay. The calculation can be much simplified by using Pratt's (1964) derivation of the cost of risk-bearing. Assuming only that the utility function is separable and that third (and higher) moments of the distribution of expenditures can be ignored, we can easily derive the relevant expression needed to calculate welfare loss. An estimate of the error from ignoring higher moments is presented in the Appendix.

Let the utility function be separable in nonmedical-care consumption and medical

care. Denote these two parts of the utility function as W and V , respectively. We want to find the *pure risk premium*, π , such that expected utility with health insurance is equal to expected utility without insurance:

$$\begin{aligned} (1) \quad & E[V(X - \pi - (1 - c)E(PM) \\ & - cPM) + W(M)] \\ & = E[V(X - pm) + W(m)] \end{aligned}$$

where X = fixed level of income; c = coinsurance rate, M = quantity of medical care consumed with insurance, m = quantity of medical care consumed without insurance, P = gross price of medical care with insurance, and p = gross price without insurance. Both sides of equation (1) involve expected values, since M and m are random variables. By definition, out-of-pocket spending for medical care is cPM with insurance and pm without insurance.

The Rand Health Insurance Experiment showed that M exceeds m because of moral hazard. Feldstein's hypothesis is that P is higher than p because of the induced price increase resulting from insurance. Finally, we assume that the insurance premium is actuarially fair: that is, premium equals $(1 - c)E(PM)$.

Taking Taylor series expansions of equation (1) (see the Appendix), we can solve for π :

$$\begin{aligned} (2) \quad \pi = & [E(pm) - E(PM)] \\ & + \frac{[E(W(M)) - E(W(m))]}{V'} \\ & + \frac{R\sigma^2}{2}. \end{aligned}$$

The first two terms on the right-hand side in equation (2), enclosed by brackets, represent the welfare loss due to higher spending on medical care with insurance. The next two terms in brackets represent the value of increased medical care to consumers. This value can be approximated by a consumer surplus triangle, assuming that the demand

curve for medical care is linear. To calculate these terms, we relied on estimates from Manning et al. (1988) of total family spending (out-of-pocket payments and insurance payments) for medical, dental, and mental health care.² In 1967 dollars, families in the free-care plan spent \$943 on average, while those in the 95-percent coinsurance plan spent \$606 on these services.

The last term is the value of risk-avoidance provided by insurance. This term is positive, and it increases as the variance of out-of-pocket spending increases or as consumers become more risk-averse. R is the Pratt-Arrow absolute risk-aversion parameter, defined as $R = -V''/V'$. Values of R can be obtained from Marquis and Holmer (1986), who found that consumers would pay up to \$634 to avoid a bet with equally probable outcomes of \$0 and -\$1,000 in 1982 dollars (i.e., an expected value of -\$500). Thus the "pure risk premium" for this bet was \$134. Changing the odds of the \$1,000 loss to 0.9 resulted in a pure risk premium of \$38 (although the expected loss is larger, the pure risk premium is lower because the loss is almost a "sure thing" at 0.9 odds). These were illustrative cases for families characterized as "healthy" and "sick," respectively. The information contained in these bets can be utilized to solve for values of R between 0.0028 and 0.0036.³

²See Manning et al. (1988 table B.6) for family spending data in nominal dollars. We converted these data to real (1967) dollars using information supplied by Will Manning that the ratio of nominal to real total spending in the 95-percent coinsurance plan was approximately 2.0.

³Values for the absolute risk-aversion parameter can be obtained by solving $\pi = R\sigma^2/2$ for the given values of $\pi = 134$ with $\sigma^2 = 250,000$ and $\pi = 38$ with $\sigma^2 = 90,000$. These are the pure risk premiums and variances of the bets offered by Marquis and Holmer (1986). The solutions are $R = 0.0011$ or $R = 0.00084$. Since Marquis and Holmer's bets were offered in 1982 dollars, R is also expressed in 1982 dollars. R values must be converted to real (1967) dollars by multiplying them by the ratio of 1982 medical care CPI (consumer price index) to 1967 medical care CPI. The medical care CPI stood at 92.5 in 1982 and 28.2 in 1967 (*Economic Report of the President*, 1989 table B-59). Thus, R values were adjusted to obtain, for example, $0.0011(92.5/28.2) = 0.0036$. Friedman (1974) estimated

Friedman (1974) estimates that $R = 0.00265$ or 0.0028.

The variance used in equation (2) should be adjusted by characteristics of families enrolled in the 95-percent coinsurance plan.⁴ Instead, the only information available to us was the total variance of out-of-pocket spending in the 95-percent coinsurance plan. This variance, which was \$25,828 in 1967 dollars, is an upper bound on the adjusted variance. However, as Joseph Newhouse (1982) and Newhouse et al. (1989) have shown, the maximum explainable variance in medical expenditures is only about 15-20 percent of total variance. Assuming that the maximum explainable percentage of variance in out-of-pocket expenditures is also 15-20 percent, the error from using our upper-bound estimate is small.

Table 1 presents the results of our calculation, expressed as the welfare loss from increasing coverage from 95-percent coinsurance to free care. Following Manning et al. (1987), we assumed that families value the last dollar of medical care spending at either 31 cents or 95 cents. To illustrate our calculations, use the first assumption and suppose that insurance causes the gross price of medical care to rise by 20 percent. Then the welfare loss per family, before considering risk-bearing, is $(943 - 606)(1 - 0.31/2) + (943)(0.2) = \474 in 1967 dollars. This estimate is converted to 1984 dollars by using the medical-care price index: $(474)(106.8/28.2) = \$1,795$.⁵ Estimates for each family were multiplied by the total number of U.S. households in 1984 headed

that R was 0.0025 or 0.00265 in 1968 dollars. These estimates were also converted to 1967 dollars. We used the value of R obtained from Marquis and Holmer's first bet in our calculations, but we also used Friedman's estimate to test the sensitivity of the calculations.

⁴The relevant variance is of family expenses, not individual expenses. To the extent that individual expenses are less than perfectly correlated, the family acts as its own risk pool, reducing the variance in medical expenses.

⁵The medical-care price index was 28.2 in 1967 and 106.8 in 1984 (*Economic Report of the President*, 1989 table B-59). The ratio of 106.8/28.2 was used to update gains and losses from insurance to 1984 dollars.

TABLE 1—THE WELFARE LOSS OF EXCESS HEALTH INSURANCE (BILLION 1984 \$)

Gross price change	Value of marginal dollar of medical spending	
	31 Cents	95 Cents
No change	61.0	33.4
10-percent increase	85.0	57.4
20-percent increase	109.3	81.7

Note: Table 1 assumes that the risk-aversion parameter is 0.0036 and that the variance of out-of-pocket spending under the 95-percent coinsurance plan is \$25,828.

by a householder under age 65: $(\$1,795)(67,506,000) = \121.2 billion.⁶

The gain from risk-bearing is $(0.0036) \times (25,828)/2 = \46.49 per family in 1967 dollars. This can be converted to \$176 per family in 1984, or \$11.9 billion for all U.S. households headed by persons under age 65. While not inconsequential, this gain is not large enough to outweigh the loss due to excess consumption of medical care. Utilizing other values for the risk-aversion parameter would proportionately change this estimate. For example, if the risk-aversion parameter were 0.0028, as implied by Friedman's study or Marquis and Holmer's lower estimate, the gain from risk-bearing would be \$137 per family in 1984.

Table 1 shows that the net welfare loss from excess health insurance is always positive. Utilizing the example discussed above, the net loss is $\$121.2 - \$11.9 = \$109.3$ billion. This is the maximum loss, which occurs if insurance induces a large increase in gross price and consumers place low marginal value on medical spending. Under assumptions of constant prices and high marginal valuation, welfare loss would fall to \$33.4 billion.

The calculations presented in Table 1 are sensitive to the assumed increase in gross price induced by insurance. Since it is prob-

lematic how much (if any) additional product quality is reflected by these price increases, the estimates should be regarded as largely illustrative. This feature of the calculations deserves further study.

In conclusion, we have provided a new estimate of the welfare loss of excess health insurance. We used the Rand Health Insurance Experiment results regarding demand effects and risk-aversion. In addition, insurance was hypothesized to affect the gross price of medical care. Accounting for these factors, we suggest that the range of welfare loss is from 33.4 to 109.3 billion dollars in 1984. While agreeing with Manning et al.'s (1987) basic conclusion, we believe that our approach refines their estimate by including the gains from risk-bearing.

Finally, the calculations presented here compare an insurance policy with a \$1,000 cap on out-of-pocket expenditure with a free-care policy. The free-care policy is found to be inferior from a welfare point of view. Since the Health Insurance Experiment did not observe families with *no* health insurance, the welfare gains or losses from a policy with a \$1,000 spending cap cannot be compared to the total absence of health insurance.

APPENDIX

Expand the left side of equation (1) by a first-order Taylor series at $X - \pi - (1 - c)E(PM) - cPM$ near $X - E(pm)$ to obtain

$$(A1a) \quad V + V'[-\pi - E(PM) + E(pm)] \\ + E[W(M)]$$

where V and all derivatives are evaluated at $X - E(pm)$. Take a second-order expansion of the right side of (1) at $X - pm$ near $X - E(pm)$:

$$(A1b) \quad V + \frac{V''\sigma^2}{2} + E[W(m)].$$

Equate (A1a) and (A1b) and solve for π . The solution is equation (2).

By using only a second-order expansion, equation (2) ignores third and higher mo-

⁶Families were excluded from the Health Insurance Experiment if the family head was eligible for Medicare at the beginning of the study or would become so by virtue of age before the end of the study. In 1984, there were 67,506,000 households headed by persons under age 65 (Bureau of the Census, 1985 table C).

ments of the distribution of out-of-pocket expenditures. We illustrate the error involved in this approximation by taking a fourth-order expansion, which yields equation (2) plus two additional terms:

$$-\frac{V^{[3]}\mu_3}{6V'} - \frac{V^{[4]}\mu_4}{24V'}$$

$V^{[n]}$ stands for the n th derivative of the utility function; μ_3 is the third moment, and μ_4 is the fourth moment of the distribution of risk. To evaluate these terms, we assume a constant absolute risk-aversion (CAR) utility function of the form

$$(A2) \quad V = -\frac{e^{-RX}}{R}$$

Take derivatives of this function:

$$V^{[3]} = R^2 e^{-RX} \quad \text{and} \quad V^{[4]} = -R^3 e^{-RX}$$

Substitute these above to get

$$-\frac{R^2\mu_3}{6} + \frac{R^3\mu_4}{24}$$

Estimates from the Rand Health Insurance experiment are $\mu_3 = \$3,420,235$ and $\mu_4 = \$1,596,966,296$ (in 1967 dollars). Using these estimates and Marquis and Holmer's (1986) value of $R = 0.0036$, we find that correcting for the third and fourth moments decreases the estimated gain from risk-bearing by \$4.29 per family in 1967 dollars, or \$16.25 in 1984 dollars. This correction is of a small order of magnitude, compared to the gain from risk-bearing of \$176 per family calculated from equation (2).

Estimates of the cost of risk-bearing that rely on the third and fourth moments of the distribution of risk are sensitive to outlier cases. Since this correction is neither large nor robust, we dropped it from equation (2).

REFERENCES

- Arrow, Kenneth J., *Essays in the Theory of Risk-Bearing*, New York: Elsevier, 1974.
- Feldstein, Martin S., "Hospital Cost Inflation: A Study in Nonprofit Price Dynamics," *American Economic Review*, December 1971, 61, 853-72.
- , "The Welfare Loss of Excess Health Insurance," *Journal of Political Economy*, March/April 1973, 81, 251-80.
- Friedman, Bernard, "Risk Aversion and the Consumer Choice of Health Insurance Option," *Review of Economics and Statistics*, May 1974, 56, 209-14.
- Manning, Willard G., Newhouse, Joseph P., Duan, Naihua, Keeler, Emmett B., Leibowitz, Arleen and Marquis, M. Susan, "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *American Economic Review*, June 1987, 77, 251-77.
- , ———, ———, ———, Benjamin, Bernadette, Leibowitz, Arleen, Marquis, M. Susan and Zwanziger, Jack, *Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment*, Publication No. R-3476-HHS, Santa Monica, CA: Rand Corporation, 1988.
- Marquis, M. Susan and Holmer, Martin R., *Choice Under Uncertainty and the Demand for Health Insurance*, Publication No. N-2516-HHS, Santa Monica, CA: Rand Corporation, 1986.
- Newhouse, Joseph P., "Is Competition the Answer?" *Journal of Health Economics*, May 1982, 1, 109-15.
- , Manning, Willard G., Keeler, Emmett B. and Sloss, Elizabeth M., "Adjusting Capitation Rates Using Objective Health Measures and Prior Utilization," *Health Care Financing Review*, Spring 1989, 10, 41-54.
- Pratt, John W., "Risk Aversion in the Small and the Large," *Econometrica*, January/April 1964, 32, 122-36.
- Bureau of the Census, "Household and Family Characteristics: March 1984," *Current Population Reports*, Series P-20, No. 398, Washington, DC: Bureau of the Census, 1985.
- Economic Report of the President*, Washington: U.S. Government Printing Office, 1989.

Reconciling Recent Estimates of the Marginal Welfare Cost of Taxation

By DON FULLERTON*

Much of welfare economics has been plagued by differences in the concepts used to measure changes in consumer welfare. One example of this problem is presented in striking clarity by three recent papers in *The American Economic Review* that calculate "marginal excess burden" for U.S. labor taxes. This note shows that almost all of the differences in results can be traced to differences in the definition of "marginal excess burden."

Each of the three papers includes several special cases, but for comparability, I concentrate on their calculations in which the uncompensated labor-supply elasticity is zero and the marginal dollar is spent on a public good that is separable in utility. Their results are summarized in Table 1. Charles Stuart (1984) builds a simple two-sector general-equilibrium model and finds in this case that "marginal excess burden" (MEB) is 7 cents. Charles Ballard, John Shoven, and John Whalley (henceforth, BSW) (1985) use a more complicated general-equilibrium model in which their "MEB" is 12 cents. Edgar Browning (1987) employs a simple partial-equilibrium model and obtains a corresponding figure of 21 cents. He concludes that "almost all of the differences in results can be traced to different assumptions about key parameter values" (p. 11).

The next row of Table 1 shows that the substitution of Stuart's parameters into Browning's formula (given in a footnote to the table) only changes Browning's measure from 21 cents to 24 cents. Thus, the large difference in published results (21 vs. 7 cents) is not due to parameters.

Since important differences in the definitions were not explained in the three papers, I state the three definitions explicitly and illustrate them diagrammatically. Rather than present new analytical results,¹ I return to the exact model of Stuart (1984), replicate his calculation, and add the other two "MEB" measures. When Stuart's "MEB" is 7 cents, Table 1 shows that the BSW measure is also 7 cents. With the same equilibrium outcome, Browning's measure is 25 cents. Thus, the results differ because the definitions differ.²

I. Three Definitions of "Marginal Excess Burden"

Stuart (1984) uses the compensating surplus (CS) of John Hicks (1943, 1954), and he subtracts the actual change in revenue (dR) to get the change in excess burden. He then divides by the actual change in revenue to get the change in excess burden per dol-

*Department of Economics, University of Virginia, Charlottesville, VA 22901. This note is a revised portion of my 1989 working paper, "If Labor is Inelastic, Are Taxes Still Distorting?" written on a grant from the Olin Foundation to the National Bureau of Economic Research. I am grateful for computer programs and substantial help from Charles Stuart, for research assistance from Joon-Kyu Park, and for helpful suggestions from Charles Ballard, George Borjas, Anne Case, Lawrence Goulder, Yolanda Henderson, James Hines, Andrew Lyon, Joram Mayshar, Hilary Sigman, Jonathan Skinner, and referees.

¹Relevant analytical results appear in Anthony Atkinson and Nicholas Stern (1974), David Wildasin (1979, 1984), Stuart (1982), Robert Triest (1988), Steven Slutsky (undated), and Joram Mayshar (1988a, b). The three *American Economic Review* papers are specifically discussed in Ballard (1987), Ingemar Hansson and Stuart (1988), Triest (1988), Shaghil Ahmed and Dean Croushore (1988), and Mayshar (1988a).

²Similar definitional differences carry over to the literature on the "marginal efficiency cost of redistribution" and may explain much of the remaining difference between results of Ballard (1988) and Browning and William Johnson (1984).

TABLE 1—DIFFERENT MEASURES OF "MARGINAL EXCESS BURDEN" FOR U.S. LABOR TAXES

Source	$(CS - dR)/dR$ (Stuart)	$(EV - dR)/dR$ (BSW)	$(EV - dR^*)/dR$ (Browning)
From the literature ^a			
Stuart (1984)	0.07	—	—
Ballard, Shoven, and Whalley (1985)	—	0.12	—
Browning (1987)	—	—	0.21
Using Browning's (1987) model ^b with Stuart's (1984) parameters	—	—	0.24
Using Stuart's (1984) model ^c			
Other measures calculated	0.07	0.07	0.25
Labor unchanged with variable wage	0.00	0.00	0.20
Labor unchanged with fixed wage	0.00	0.00	0.24

Notes: CS = compensating surplus; EV = equivalent variation; dR = change in actual revenue using average tax rates; dR^* = change in revenue along curve compensated to new utility, using marginal rates.

^aEstimates from the literature all employ a zero uncompensated-labor-supply elasticity, where additional revenue is spent on government consumption, but they differ on other assumptions.

^bBrowning (1987) calculates the area between a fixed gross wage rate and an approximately linear compensated labor supply curve and shows that

$$\frac{EV - dR^*}{dR} = \left(\frac{m + 0.5}{1 - m} \right) \eta \left(\frac{dm}{dt} \right)$$

where $\eta = 0.2$ is the compensated labor supply elasticity, $m = 0.43$ is the marginal tax rate, and $t = 0.31$ is the average tax rate. The tax change is assumed to maintain progressivity, so $dm/dt = m/t = 1.39$. With $dm = 0.01$, this equation yields 21 cents for Browning's (1987) "MEB" as shown. With Stuart's (1984) parameters, $\eta = 0.2$, $m = 0.427$, $t = 0.273$, and $dm/dt = 1.564$, so this equation yields 24 cents.

^cThe text outlines Stuart's (1984) model, but more detail is in his paper. The text also describes why it is not possible to use the same assumptions in the model of Ballard, Shoven, and Whalley (1985).

lar of additional revenue, so his "MEB" is $(CS - dR)/dR$. BSW (1985) use the equivalent variation (EV) of the tax change (in old cum-tax prices), subtract the actual change in revenue, and calculate $(EV - dR)/dR$.³ Browning (1987) also uses the equivalent variation,⁴ but he subtracts the change in revenue along the compensated-labor-supply curve (dR^*). Per dollar of actual additional revenue, his "MEB" is $(EV - dR^*)/dR$.

³They never state this expression explicitly, however, so others misunderstand what was calculated. Triest (1988) assumes that this MEB is measured in pretax prices, when it is actually in prices of the original cum-tax equilibrium. Mayshar (1988b) introduces $EV - dR$ as a "new" measure, not knowing that this is the measure of BSW (1985).

⁴Browning (1987) uses the compensating variation (CV) in his diagram for the other case he considers, but he uses the equivalent variation for the case of interest here, where actual labor supply does not change (see his footnote 24).

It might appear that the measure of BSW (1985) is similar to that of Browning (1987), because both use the equivalent variation, while Stuart (1984) appears to differ by use of the compensating surplus. However, results below show that Stuart's measure is always very close to that of BSW. Indeed, for a truly marginal change in tax and hence in equilibrium prices, the measures of welfare (CS or EV) are equivalent.⁵ Thus, the

⁵The choice of welfare measure is discussed in Peter Diamond and Daniel McFadden (1974), John Kay (1980), Elisha Pazner and Efraim Sadka (1980), Alan Auerbach and Harvey Rosen (1980), and Mayshar (1988b). This choice matters for *total* excess burden, where no-tax prices are very different from cum-tax prices, but not for "*marginal* excess burden." Mayshar (1988b) proves that the EV and CV are equal at the margin, and Eugene Silberberg (1978 pp. 257-9) can be used to show equivalence with the CS measure. For small but discrete changes here, my results confirm Stuart (1984 footnote 5) that "substitution of these alternative measures [EV, CV, or CS] was found to affect the results only at the third significant digit."

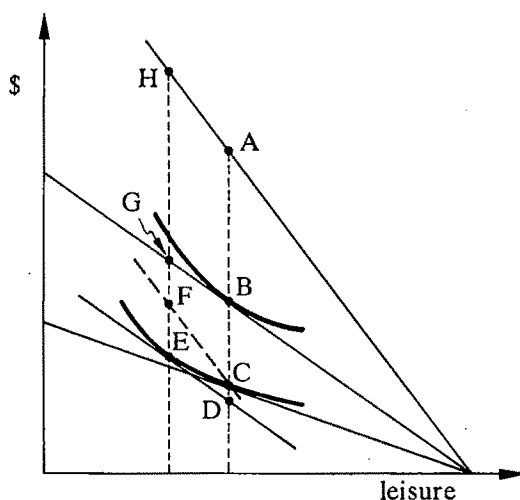


FIGURE 1. A CHANGE IN WAGE TAX WITH NO CHANGE IN ACTUAL LABOR

important difference is the revenue figure subtracted (dR or dR^*).

Figure 1 illustrates these definitions for the simple case in which a change in tax leaves actual labor unchanged. The uppermost budget line reflects a fixed gross wage, and the middle budget line reflects an initial proportional wage tax. The consumer chooses point B, and initial revenue is AB. When the wage tax is increased, the consumer chooses point C. The compensating surplus is the amount of the numeraire needed to get back to the old utility, without reoptimizing, the distance from C back to B. Since the actual revenue change is also BC, Stuart's "MEB" is $(BC - BC)/BC = 0$.

EV is the number of dollars taken away at old cum-tax prices that would reduce utility by an amount "equivalent" to the tax, distance BD. Therefore, BSW (1985) measure $(BD - BC)/BC = CD/BC$. This "MEB" may be larger than 0 for a discrete change in Figure 1, but it is 0 for a truly marginal change in tax. It is 0 regardless of the original tax rate or compensated-labor-supply elasticity.

In contrast, Browning (1987) follows Alan Auerbach (1985 p. 72) by subtracting the change in revenue if the consumer were held to the new utility level (dR^*). The new

actual revenue is AC, which equals FH by construction. At the old tax rate, if held to the new utility, the consumer chooses point E. With this labor supply, revenue at the old rate would be GH. Thus, dR^* is the difference, FG, and Browning's "MEB" is $(EG - FG)/BC = EF/BC$. This measure of distortion can be large even if actual labor is unchanged.

Given these large conceptual differences, what are the pros and cons of each definition? Each measure has an appropriate use, but each can easily be misinterpreted. All three papers deal ultimately with the question of whether utility would rise if a marginal increase in the wage tax is used to fund a public project that is separable in utility.⁶ In addressing this question earlier, Atkinson and Stern (1974) isolated two modifications to Paul Samuelson's (1954) rule that the sum of the marginal rates of substitution (ΣMRS) should equal the marginal rate of transformation (MRT). First, if the tax is not lump sum, the "distortionary effect" adds to the relevant costs. Second, in what they call a "revenue effect," changes in after-tax income affect consumer choices and thus affect tax revenue. If leisure is normal, an increased wage tax has an income effect that increases labor, increases revenue from the preexisting tax, and thus makes the project easier to fund. In this case, the "revenue effect" works in the opposite direction as the "distortionary effect."

In the special case in which actual labor does not change, the "distortionary effect" and the "revenue effect" exactly offset each other, so the project increases utility if ΣMRS exceeds the dollar cost of the project. This is the basis for the subsequent

⁶Using Ballard's (1987) terminology, it is a "balanced-budget" analysis. Stuart (1984) also performs a "differential" analysis, in which the additional tax revenue is returned to consumers in a lump sum, but this simulation has no counterpart in BSW (1985) or Browning (1987). Also, Browning (1987) includes a case in which the actual change in labor is the same as the compensated change in labor, but this has no counterpart in Stuart (1984) or BSW (1985). The only case that appears in all three papers is the balanced-budget spending on a separable public good.

finding in the literature that the marginal cost of funds (MCF) is 1.0 when labor does not change.⁷

Since Browning (1987) compares the wage tax to a lump-sum tax, he measures only the distortionary effect. This measure can appropriately be used to compare one tax to another, as it is a correct measure of distortions. However, Browning leaves the incorrect impression that the cost-benefit analyst can use 1 plus his "MEB" as the marginal cost of funds. His measure is not generally enough information to evaluate the public project, because the decision rule should be modified by both the "distortionary effect" and the "revenue effect."⁸

As can be seen from above formulas, Stuart (1984) and BSW (1985) essentially define "marginal excess burden" as the MCF minus 1. With this definition, the cost-benefit analyst can appropriately use $1 + \text{MEB}$ as the MCF. Since the "distortionary effect" and the "revenue effect" exactly offset each other in this special case, however, this "MEB" is zero. Such terminology may leave the incorrect impression that the tax is not distorting. The wage tax is distorting in the usual sense, in that it leaves consumers worse off than a lump-sum tax.

One might think that a lump-sum tax would have no excess burden. Yet for Stuart and BSW, with a preexisting wage tax, a lump-sum increment has "MEB" < 0 . In Figure 2, the initial point B is the same as in Figure 1, but a small lump-sum tax moves the consumer to point C, where the new revenue CG exceeds old revenue FG by the amount CF. Stuart's "MEB" of this lump-sum tax is $(CE - CF)/CF$, which is negative.

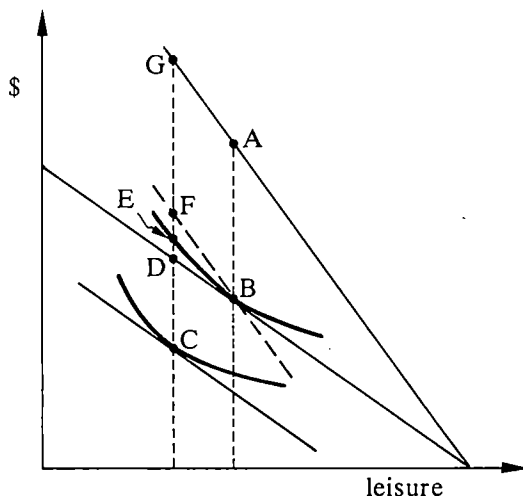


FIGURE 2. A LUMP-SUM TAX WITH A PREEXISTING WAGE TAX

The BSW measure is $(CD - CF)/CF$, which also is negative. Browning's "MEB" is $(CD - CD)/CF = 0$. The lump-sum tax has no distortionary effect, but it has a revenue effect because it reduces income and thus causes an increase in labor subject to the preexisting wage tax.

Thus, all three definitions have problems. My own view is that no measure of "marginal excess burden" is necessary. Instead, the marginal cost of funds is enough information. It can be used to compare one tax to another, like Browning's measure, and it can be used to evaluate the public project, like the other two measures. Why define "MEB" as $\text{MCF} - 1$? The marginal cost of funds is the relevant concept in any case, so the cost-benefit analyst must simply add back the 1 subtracted.

II. Comparable Calculations

To calculate his "MEB" in a general equilibrium model, Stuart (1984) first assumes that a representative consumer has an endowment of labor that can be supplied to the market sector or used in production at home. Production in each sector is Cobb-Douglas. Capital stocks are immobile, so labor has a diminishing marginal product

⁷This result is shown in, for example, Stuart (1982), Wildasin (1984), and Triest (1988). Geometrically, in Figure 1, the $\text{MCF} \equiv CS/dR = BC/BC = 1$. Equivalently, $\text{MCF} \equiv EV/dR = BD/BC$, which approaches 1.0 as the tax change becomes small.

⁸This problem is recognized by Browning (personal communication) and discussed by Hansson and Stuart (1988). It does not apply to the other case in Browning (1987), in which the public funds provide benefits that return taxpayers to the original indifference curve, because then the actual change in labor is the compensated change.

in each sector. The consumer uses total income to maximize

$$(1) \quad U = [\alpha(\bar{Y}_1)^{-\rho} + (1-\alpha)(\bar{Y}_2 - \delta)^{-\rho}]^{-1/\rho}$$

where δ is a "minimum required purchase," α is a share parameter, and the elasticity of substitution is $\sigma = 1/(1 + \rho)$. The nonmarket sector has no tax, so labor earns its marginal product, and consumption \bar{Y}_2 equals production Y_2 . In the market sector, labor earns the equilibrium wage times 1 minus the marginal tax rate, and consumption \bar{Y}_1 is less than production Y_1 by the amount of tax paid (the wage bill times the average tax rate). The government spends marginal revenue on the market good.

Stuart (1984) provides detail on the model, and elsewhere (Fullerton, 1989), I further discuss the derivation of parameters and additional steps needed to calculate the other welfare measures. Using Stuart's (1984) model and parameters, I simulate a 1-percent increase in the marginal tax rate, m . Stuart's (1984) measure of "marginal excess burden" is 7 cents, just as in published results. Table 1 shows that the measure of BSW (1985) also rounds to 7 cents. Using the same equilibrium outcome, however, Browning's (1987) measure is 25 cents. Thus, the results differ because the definitions differ.

When labor did not change in Figure 1, Stuart's (1984) measure was shown to provide zero excess burden; so why does he get 7 cents? The reason is that this simulation with a zero uncompensated labor elasticity does not lead to zero change in actual labor supply. Ballard (1987) points out that this change in the progressive tax structure effectively changes the "virtual" income of the consumer. The change in the net wage by itself would not change labor supply, but the change in virtual income does. The condition for the marginal cost of funds to be 1.0 is not that the uncompensated elasticity is zero, but that actual labor does not change. Thus, Stuart's (1984) "MEB" ($= \text{MCF} - 1$) will not be zero whenever any aspect of the reform causes a response in the quantity of labor.

To set his parameters, Stuart (1984) differentiates labor supply with respect to the net wage and imposes an uncompensated elasticity of 0, a compensated elasticity of 0.2, and an initial MRS of 1. Together with his data, these three conditions determine the three parameters (δ , α , and ρ). This procedure is consistent with the definition of an elasticity, since the differentiation varies only the net wage. As an alternative, I search a three-dimensional grid for values of δ , α , and ρ where the compensated elasticity is still 0.2 and the MRS is 1, but where actual labor does not change in this particular simulation. In this case, Table 1 shows that Stuart's "MEB" is zero.⁹

Using these new parameters, the measure of BSW also is zero; so why do they get 12 cents? As pointed out by Ballard (1987), they have other taxes that introduce second-best effects. The simulation here demonstrates that their definition would yield a zero "marginal excess burden" in the case with no other taxes and an unchanged equilibrium supply of labor.¹⁰

The model of BSW (1985) is not used here to calculate the three measures. One reason is that this model has 12 different consumer groups with different marginal tax rates and elasticities. Another reason is that the calculation of dR^* would be extremely difficult with many tax instruments. It would require the compensated demand for each commodity, the sales tax on each compensated quantity, and all factor taxes on producers at those quantities.

Finally, I impose on Stuart's (1984) model the condition that production is linear, so the gross wage is constant in general equi-

⁹These new parameters are not preferred to those of Stuart (1984). Indeed, they imply that the uncompensated elasticity is (slightly) negative. They are used here only to illustrate the important conceptual point that Stuart's measure is zero when actual labor does not change. This point was not clear in Stuart's (1984) paper.

¹⁰The point of this note does not arise when their model is used in a revenue-neutral reform, as in every previous application, because the EV measures the change in welfare with no subtraction for any change in revenue.

librium. Stuart's model then reduces exactly to Browning's (1987) model (where the wage was constant by assumption). An additional grid search is performed to impose the three conditions discussed above (the compensated labor supply elasticity is 0.2, the initial MRS is 1, and actual labor is unchanged in this simulation). Table 1 shows that Browning's "MEB" is 24 cents, while the other measures are zero. Browning's fixed wage therefore raises his own "MEB" by 20 percent (24 cents vs. 20 cents), relative to a comparable model with a varying wage.

III. Conclusion

With only the "distortionary effect," Browning's (1987) "marginal excess burden" is a familiar concept. It is the marginal analogue of total excess burden, defined as the welfare difference between a distorting tax and a lump-sum tax. However, it is not enough information to set public spending. With the addition of the "revenue effect," the measure of Stuart (1984) or of BSW (1985) does provide enough information to decide on a project that is separable in utility. However, it can be zero for a distorting tax. It is defined as the marginal cost of funds minus 1, but there is no need for a concept other than the MCF itself.

The marginal cost of funds can be used to compare the distorting effects of two different tax changes, because the marginal dollar always has the same revenue effect. For a given tax, the MCF can be compared to the benefits of a public project. In general, the MCF does depend on elasticities and tax rates in the model. If the project is not separable in utility, it also depends on the assumed effect of the public project on labor supply. For a marginal dollar of revenue, however, it does not depend on the definition of consumer welfare.

REFERENCES

- Ahmed, Shaghil and Croushore, Dean D., "Substitution Effects and the Marginal Welfare Cost of Taxation," mimeo, Pennsylvania State University, 1988.
- Atkinson, Anthony B. and Stern, N. H., "Pigou, Taxation, and Public Goods," *Review of Economic Studies*, January 1974, 41, 119-28.
- Auerbach, Alan J., "The Theory of Excess Burden and Optimal Taxation," in A. J. Auerbach and M. Feldstein, eds., *Handbook of Public Economics*, Vol. 1, Amsterdam: Elsevier, 1985, 61-127.
- _____, and Rosen, Harvey S., "Will the Real Excess Burden Please Stand Up? (or, Seven Measures in Search of a Concept)," National Bureau of Economic Research (Cambridge, MA) Working Paper No. 495, 1980.
- Ballard, Charles L., "Marginal Efficiency Cost Calculations: Differential Analysis vs. Balanced-Budget Analysis," mimeo, Michigan State University, 1987.
- _____, "The Marginal Efficiency Cost of Redistribution," *American Economic Review*, December 1988, 78, 1019-33.
- _____, Shoven, John B., and Whalley, John, "General Equilibrium Computations of the Marginal Welfare Costs of Taxes in the United States," *American Economic Review*, March, 1985, 75, 128-38.
- Browning, Edgar K., "On the Marginal Welfare Cost of Taxation," *American Economic Review*, March 1987, 77, 11-23.
- _____, and Johnson, William R., "The Trade-Off Between Equality and Efficiency," *Journal of Political Economy*, April 1984, 92, 175-203.
- Diamond, Peter A., and McFadden, Daniel L., "Some Uses of the Expenditure Function in Public Economics," *Journal of Public Economics*, February 1974, 3, 3-21.
- Fullerton, Don, "If Labor is Inelastic, Are Taxes Still Distorting?" National Bureau of Economic Research (Cambridge, MA) Working Paper No. 2810, 1989.
- Hansson, Ingemar and Stuart, Charles, "Measures of Costs of Taxation: Comment on Browning and Others," mimeo, University of California, Santa Barbara, 1988.
- Hicks, John R., "The Four Consumer's Surpluses," *Review of Economic Studies*, Winter 1943, 11, 31-41.
- _____, *A Revision of Demand Theory*, Oxford: Oxford University Press, 1954.
- Kay, John A., "The Deadweight Loss from a

- Tax System," *Journal of Public Economics*, February 1980, 13, 111-9.
- Mayshar, Joram, (1988a) "A Note on Measuring the Marginal Cost of Taxation," Paper No. 175, The Hebrew University of Jerusalem, 1988.
- _____, (1988b) "On Measures of Excess Burden and Their Application," Paper No. 199, The Hebrew University of Jerusalem, 1988.
- Pazner, Elisha A. and Sadka, Efraim, "Excess-Burden and Economic Surplus as Consistent Welfare Indicators," *Public Finance / Finances Publiques*, 1980, 35(3), 439-49.
- Samuelson, Paul A., "The Pure Theory of Public Expenditure," *Review of Economics and Statistics*, November 1954, 36, 387-9.
- Silberberg, Eugene, *The Structure of Economics: A Mathematical Analysis*, New York: McGraw-Hill, 1978.
- Slutsky, Steven, "Undersupply and the Untaxed Commodity: A Note on Atkinson and Stern," mimeo, Rice University, undated.
- Stuart, Charles, "Measures of the Welfare Costs of Taxation," mimeo, University of California, Santa Barbara, 1982.
- _____, "Welfare Costs per Dollar of Additional Tax Revenue in the United States," *American Economic Review*, June 1984, 74, 352-62.
- Triest, Robert K., "The Relationship Between the Marginal Cost of Public Funds and Marginal Excess Burden," mimeo, University of California, Davis, 1988.
- Wildasin, David E., "Public Good Provision with Optimal and Non-optimal Commodity Taxation: The Single Consumer Case," *Economics Letters*, 1979, 4(1), 59-64.
- _____, "On Public Good Provision with Distortionary Taxation," *Economic Inquiry*, April 1984, 22, 227-43.

Does Student Aid Affect College Enrollment? New Evidence on a Persistent Controversy

By MICHAEL S. MCPHERSON AND MORTON OWEN SCHAPIRO*

Certainly no aspect of the evaluation of federal student aid has attracted more attention than the question of its impact on enrollment levels and patterns. Although it is important to note that affecting enrollment is not the whole justification for student aid, the aim of promoting the enrollment of less-affluent students has been central to the case for federal student aid throughout its history.¹ Despite quite substantial empirical efforts, the issue of the size—and even the existence—of these enrollment effects remains unsettled. A major difficulty is that controlled econometric studies of student behavior, the best of which have relied on cross-sectional data on individuals, lead us to expect substantial effects of student aid, but these effects have been hard to discern in the historical time series.

Before the introduction of the Basic Educational Opportunity Grant program in 1974 (later renamed Pell), total federal spending on need-based grants to undergraduate students amounted to less than a third of a billion in 1982 dollars and accounted for less than 3 percent of total tuition revenue. By 1980, need-based federal grants were over 3.5 billion 1982 dollars, and the Pell program accounted for more than 80 per-

cent of that total. These federal grants to students amounted to 29 percent of total tuition revenue of U.S. colleges and universities in 1980. Over the same period, federally subsidized loans for college grew from around \$3 billion to over \$5.5 billion 1982 dollars. Yet, despite this dramatic change in financing, enrollment rates in 1980 were, if anything, slightly below those earlier in the decade. After 1980, the grant programs experienced little real growth, subsidized loans continued to increase, and enrollment rates remained fairly steady.² The relative stability of overall enrollment rates in the light of substantial fluctuations in federal spending on student aid is an empirical puzzle and a challenge for public policy.

This paper reports on a disaggregated econometric analysis of time-series evidence on U.S. higher-education enrollments and net costs over the 1974–1984 period. Section I contains a brief literature review. Section II presents our regression results and compares them to findings from cross-sectional studies. Section III contains a summary and conclusions.

I. The Literature

A great many studies over the years have attempted to estimate the impact of price or net cost of education on students' postsecondary education decisions.³ A minority of

*Department of Economics, Fernald House, Williams College, Williamstown, MA 01267. The authors thank three anonymous referees and the participants in the Economics Seminar at Williams College and in the National Bureau of Economic Research Conference on the Economics of Higher Education. Barry Bosworth, Lee Hansen, Frank Lichtenberg, and Charles Manski also contributed helpful comments. Michael P. O'Malley, Diedre Goodwin, and Mary Skinner provided excellent research assistance. Research support from the Andrew W. Mellon Foundation, the Spencer Foundation, and the Teagle Foundation is acknowledged with gratitude.

¹A broader framework of goals for federal student aid is suggested in McPherson (1988).

²More-detailed examination of the federal student aid programs and their changing funding levels is provided in McPherson and Schapiro (1990). Enrollment data are from the Bureau of Labor Statistics, Current Population Survey. Student aid data are from Donald Gillespie and Nancy Carlson (1983) and Gwendolyn Lewis (1988).

³A number of able surveys of this literature exist. A recent one, which provides references to many of its predecessors, is Larry Leslie and Paul Brinkman (1987); see also Leslie and Brinkman (1988).

those studies have tried to measure specifically the effect of student aid on enrollment decisions, with the rest focusing on the impact of tuition price. Although the studies differ widely in data sources and estimation techniques, they tend to agree on two main points. First, student decisions to enroll in college respond positively and nontrivially to price cuts or aid increases. Second, decisions about where to attend school also respond nontrivially to changes in the relative prices of schooling alternatives.

Perhaps the best and most influential of these studies is Charles Manski and David Wise (1983). According to a simulation based on their estimates of college choices of a sample of 1972 high school graduates, the Pell grant program as it existed in 1979–1980 should have left enrollments 21 percent higher than they would have been without Pell, with the increases heavily concentrated at two-year colleges and among students from lower-income families. The predicted response by income group varies greatly: there is a 59-percent enrollment increase for low-income students, a 12-percent increase for middle-income students, and only a 3-percent increase for upper-income students.⁴

Further econometric support for the claim that financial aid influences enrollment is provided by the many studies that estimate the effect of tuition variations on enrollment behavior. Although changes in grant awards may have somewhat different effects

on enrollments from tuition changes that have equivalent impacts on net price, the size of those effects and their variation across income classes should be similar. It is therefore reassuring to note that most studies of enrollment demand find significant positive effects of tuition reductions on enrollment levels and find that the enrollment effects (in percentage terms) are larger for lower-income students.⁵ Leslie and Brinkman (1987) find that a consensus of the studies they survey puts the effects of a price cut of \$100 (1982–1983 academic-year dollars) on national enrollment of 18–24 year-olds at about 1.8 percent. On the assumption that a price cut and a grant increase of the same magnitude have equal effects, the Pell program as it existed in 1979 should have boosted total enrollment by approximately 10–15 percent, compared to what enrollments would have been in that year without the program.⁶ This is roughly comparable to the findings of studies like Manski and Wise (1983) that try to measure the effect of grant aid directly.

These econometric findings create an expectation that it should be possible to detect effects of changing student-aid policy in the national time-series data. However, a number of observers have noted the absence of any obvious change in national enrollment trends in response to changes in federal student-aid policies and funding levels, (e.g., Robert Zemsky [1988] and Leslie and Brinkman [1988]). Moreover, some of the most careful econometric studies (in-

⁴Several key features of the Manski and Wise (1983) findings on the access effects of Pell grants are corroborated by other studies. Estimates developed by Leslie and Brinkman (1988) from their analysis of seven econometric studies (including that of Manski and Wise) suggest that the Pell program as it existed at the end of the 1970's should have raised lower-income enrollment by between 20 percent and 40 percent, implying an increase in total enrollment of approximately 10–20 percent. They point out that these results indicate that roughly between 500,000 and 1 million low-income students and approximately 400,000 middle-income students are enrolled in college because of grant aid. The midpoint of the total of these figures is slightly over 1 million students, approximately 16 percent of all full-time students. For more detailed discussion of the relevant literature, see McPherson and Schapiro (1990).

⁵See Leslie and Brinkman (1987) for a comprehensive survey. For an analytically oriented survey that examines the relation between income levels and price responsiveness of enrollment, see McPherson (1978).

⁶This assumes an average Pell award of about \$1,000 1979 dollars per recipient and that about half of freshmen should have been eligible for Pell awards under 1979 rules. Actual enrollment rates of high school graduates grew by about 5 percent from 1973 to 1979. Thus, these estimates imply that, in the absence of the Pell program, the enrollment rate would have fallen by between 5 percent and 10 percent over this time period. One potential explanation for such a decline is the sharp drop in the earnings differential between college and high school graduates over this period. See Lawrence Katz and Kevin Murphy (1990).

cluding Manski and Wise [1983]) rely on data collected before the introduction of the Pell program in 1974. Inferences from these estimates to behavior in the post-Pell period may be suspect if the introduction of a major new federal program changed the structure of the aid-enrollment relationship.

W. Lee Hansen (1983), in a highly influential study, has suggested that it is useful to look at relative enrollment rates of more- and less-affluent students in gauging the impact of federal student aid, on the grounds that changes over time in federal student aid are the most obvious factor that should affect time-series changes in the enrollment behavior of these two groups differentially. He used Current Population Survey (CPS) data in examining enrollment rates for students from families with dependents aged 18-24 for two time periods: 1971-1972 and 1978-1979. He then calculated the ratio of the enrollment rates of below- to above-median-income families in the two periods and found that the ratios declined for whites, blacks, men, and women. When a weighted average was taken for whites and blacks and for men and women, the ratios again fell between the two periods.

The conclusion from this study is well known among researchers and policy makers: "These data force one to conclude that the greater availability of student financial aid, targeted largely toward students from below-median-income families, did little, if anything, to increase access. The results certainly do not accord with expectations that access would increase for lower-income dependents relative to higher-income dependents." (Hansen, 1983 p. 93)

There are some obvious limitations in interpreting this kind of snapshot comparison at two points in time. First, year-to-year fluctuations may obscure underlying trends, so that increasing the number of years in the comparison is helpful. Second, controlling for variation in other factors that affect the demand for enrollment is not possible with this method. Such factors as overall economic conditions, changes in rates of return to higher education, and changes in

opportunity costs of college enrollment (as produced, for example, by changes in the draft law) may influence the comparison if these factors affect different income groups differently. Finally, this kind of comparison is not responsive to changes over time in the targeting of student aid. During the 1970's, the total amount of federal student aid not only increased substantially, but also changed significantly in its distribution. A larger fraction of available aid was targeted at middle- and upper-income students in the late 1970's, tending to obscure any effect on differential enrollment rates that might have occurred.

II. Analysis

Summarizing the above discussion, researchers have found significant econometric evidence of a rather large enrollment response to differences in student aid. However, despite substantial variation in aid over time, enrollment responses are not readily detected in national time-series data. Earlier analysts have, however, failed to subject time-series data from the post-Pell era to econometric analysis.⁷

Our analysis is based on enrollment, tuition, and financial-aid data for population subgroups over the 1974-1984 period. The enrollment data are from the Current Population Survey; the tuition and financial aid data are from an annual survey of college freshmen, The American Freshman survey.⁸

⁷Several time-series econometric studies of enrollment demand exist, but these predate the introduction of the Basic Grants program. See John Hight (1975), Robert Campbell and Barry Siegel (1967), and Stephen Hoenack and William Weiler (1975). Results from these studies are, on the whole, comparable to the findings of cross-sectional econometric studies.

⁸The data on tuition, student aid, and income in the American Freshman Survey are self-reported by students. No doubt this self-reporting introduces measurement error in these variables. Nevertheless, we use these data for several reasons. First, they are the only consistently reported annual data on net costs and income. Second, there is no reason to expect the biases in student reporting of income and costs to vary systematically over time. Hence, while the data may be inaccurate as estimates of these values in any particular year, their variation over time should be more

TABLE 1—DESCRIPTIVE STATISTICS

Statistic	N	Mean	SD	Minimum	Maximum
All institutions:					
Enrollment rate	66	0.427	0.117	0.246	0.672
Net cost (NETCOST)	66	3.056	0.693	1.974	4.623
Private institutions:					
Enrollment rate	60	0.111	0.051	0.037	0.228
Net cost (NETCSTPR)	60	4.126	0.949	2.622	6.336
Public institutions:					
Enrollment rate	60	0.318	0.070	0.194	0.461
Net cost (NETCSTPU)	60	2.460	0.433	1.657	3.197

Note: Net costs are reported in thousands of dollars.

The individual data points in our regressions are an enrollment rate and an average net cost for a particular population subgroup (e.g., white women, incomes below \$10,000) in a particular year.⁹ We employ three such data sets: one for public institutions, one for private institutions, and one that averages over public and private institutions. Investigations with the data suggest that small samples in the CPS data for blacks and other races preclude time-series analysis at the level of disaggregation we employ. Therefore, the results we report here are limited to whites only. In the regressions that report on enrollments at public and private institutions separately, we are forced to exclude data for 1980, because mistakes made by the Bureau of the Census in coding the 1980 CPS make it impossible to distinguish public from private enroll-

ment. Thus, regressions using the combined data set are based on 66 observations (three income groups, two genders, and 11 years). Regressions for public and for private institutions have 60 observations (three income groups, two genders, and 10 years). Table 1 contains descriptive statistics.

Table 2 presents regression results in which enrollment rates averaged across public and private institutions are explained by time-series changes in net cost and other variables. Given the nature of the data set, heteroskedasticity is a natural worry. Therefore, for all of the regression results that follow, estimated asymptotic covariance matrices were computed under the assumption of heteroskedasticity in order to calculate the standard errors.¹⁰ These adjusted standard errors were used in all tests of significance. The regression equation includes a time trend along with a dummy variable for gender (FEMALE: 1 for females and 0 for males) and dummy variables for the medium-income group (MED: income between \$10,000 and \$30,000 in 1978 dollars) and for the high-income group (HIGH: income over \$30,000). In addition, the equation includes terms that interact income with the net-cost variable, the gender dummy, and the time trend (TIME).

reliable. Finally, we know of no reason why any systematic biases in these variables should be correlated with variations in the dependent variable (the enrollment rate). Note that the dependent variable is obtained from a data set that is collected separately from these independent variables.

⁹We define "net cost" as the difference between tuition (the "sticker price") and the subsidy value of student aid. Net cost is measured in thousands of 1978–1979 dollars. The subsidy value is calculated on the assumption that subsidized loans obtained by students from the federal government provide a 50-percent subsidy. Several attempts to estimate the present value of student-loan repayment streams put the implicit subsidy at approximately half the face value of the loan. See Barry Bosworth et al. (1987) and Arthur Hauptman (1985).

¹⁰For the derivation of this technique, see Halbert White (1980). The correction of the standard errors does not produce major changes from the results obtained without the correction.

TABLE 2—COMBINED SAMPLE; DEPENDENT VARIABLE = ENROLLMENT RATE

Variable	Parameter estimate	SE	<i>t</i> for H_0 : parameter = 0
Intercept	0.461	0.050	9.29**
NETCOST	-0.068	0.023	-2.95**
TIME ($\times 10^{-3}$)	-3.645	1.755	-2.08*
FEMALE	0.049	0.009	5.56**
MED	-0.143	0.063	-2.25*
HIGH	-0.210	0.073	-2.86**
NETCSTHI	0.155	0.028	5.53**
NETCSTMED	0.091	0.027	3.36**
TIMEHI ($\times 10^{-3}$)	-3.005	2.773	-1.08
TIMEMED ($\times 10^{-3}$)	2.917	2.096	1.39
FEMHI	-0.001	0.013	-0.09
FEMMED	-0.000	0.011	-0.02
Test			χ^2
NETCOST + NETCSTHI = 0			32.57**
NETCOST + NETCSTMED = 0			2.85 ^a
TIME + TIMEHI = 0			9.59**
TIME + TIMEMED = 0			0.40
FEMALE + FEMHI = 0			22.86**
FEMALE + FEMMED = 0			53.02**

Notes: $N = 66$; mean enrollment rate = 0.427; root MSE = 0.023; CV = 5.3; $R^2 = 0.97$; adjusted $R^2 = 0.96$.

^aSignificant at $P < 0.10$ level; *significant at $P < 0.05$ level; **significant at $P < 0.01$ level.

NETCSTHI interacts NETCOST with the dummy variable representing high income. NETCSTMED interacts NETCOST with the medium-income dummy variable. TIMEHI and TIMEMED interact TIME with the income dummies, while FEMHI and FEMMED interact FEMALE with the income dummies.

We have the following expectations about the signs of the coefficients. The NETCOST coefficient, which measures the responsiveness of enrollment to net cost for the low-income group, should be negative. The coefficient on NETCSTMED measures the difference between the responsiveness of low- and middle-income students' enrollment to changes in net cost. Cross-sectional studies generally indicate that higher-income students are less responsive to price than are lower-income students. We therefore expect the coefficient on NETCSTMED to be positive, muting the negative effect of net cost on enrollment relative to that of lower-income students. For the same

reason, we expect the coefficient on NETCSTHI to be positive (and larger than that on NETCSTMED).

As Table 2 shows, all the estimated coefficients on these net-cost variables are significant with the expected sign.¹¹ Increases in net cost lead to lower enrollment for the low-income group, and the interaction effects are positive and significant, showing that this effect is smaller for middle- and upper-income students. In fact, the coefficients on the net-cost \times income interaction terms are larger in absolute value than the

¹¹In order to ensure that all predicted values lie within the unit interval, we ran regressions using a logistic transformation of the dependent variable. These results did not differ substantively from those reported below. Further, an examination of the residuals from the regressions we present below did not provide any evidence of autocorrelation within particular economic or demographic groups. (Note that standard tests for autocorrelation are inappropriate, given the panel nature of the data set.)

coefficient on net cost, implying that the predicted effect of net cost on enrollment in this equation is positive (and statistically significant, as the chi-square tests show) for middle- and upper-income students.¹² It is possible that this unexpected result for more-affluent students is explained by a supply, rather than a demand, effect: a positive relationship between enrollment and net cost may come about because (particularly in the 1980's) a strong demand among middle- and upper-income students for higher education has caused colleges and universities to raise their prices.¹³

The negative coefficient on net cost implies that for lower-income students a \$100 net-cost increase results in an enrollment decline of about 0.68 percentage points, which is about a 2.2-percent decline. We noted above that Leslie and Brinkman (1987) find a consensus in the literature that a \$100 increase in net cost reduces enrollment rates by 1.8 percent. Converting our estimates in 1978–1979 dollars to the 1982–1983 equivalent relied on by Leslie and Brinkman (1987), we find that a \$100 cost increase results in a 1.6-percent enrollment decline for low-income students. The Leslie and Brinkman (1987) figure is in effect averaged over all income groups. As noted earlier, most studies find higher price responsiveness among lower-income students. Manski and Wise's (1983) results, for

example, suggest that a \$100 net-cost increase for low-income students (in 1979 dollars) leads to a 4.9-percent decline in enrollment.¹⁴ The result here, while lower than the estimate of Manski and Wise, seems broadly consistent with typical cross-sectional findings. The important point is that our econometrically controlled time-series analysis supports the view that changes in costs lead to changes in enrollment for low-income students.

We turn next to the coefficients relating to gender and to the time trend. The coefficient on the FEMALE variable indicates that, over the 1974–1984 period, the enrollment rate for women tended to be about 5 percentage points higher than that for men. The fact that the variables interacting FEMALE with income are close to zero and statistically nonsignificant indicates that this gender effect is constant across income groups (chi-square values show that the net effect of the FEMALE variable on enrollment is positive and significant for all three income groups). The time trend is negative and significant for the low-income group, suggesting a tendency for the enrollment propensity for that group to fall over time, but the coefficient is quite small, with the estimated rate of decline being just 0.36 percentage points per year. There is no significant time trend for middle-income students, but there is a significant negative time trend of 0.66 percentage points per year for high-income students. The negative time trends noted here and below may indicate the presence of unmeasured variables tending to lower enrollment propensities over time.

Tables 3 and 4 examine private enrollment and public enrollment separately. This breakdown is particularly important because of a potential problem with the endogeneity of the price variable in the equations that average over sectors: if, for example, the number of students choosing to attend private institutions (which are generally higher priced) rises, this choice

¹²The values of the intercept and the MED and HIGH dummies imply that for all three income groups the intercept terms are positive but are a declining function of income. This may seem surprising, since we expect enrollment rates to vary positively with income. However, the presence of a negative net cost effect for the low-income group, coupled with positive effects for the other income groups, implies that predicted levels of enrollment evaluated at means in fact increase with income.

¹³Because enrollment rates are substantially higher for middle- and high-income students than for low-income students and because these students generally pay higher net costs than do low-income students, it is more plausible to expect a supply response to the behavior of middle- and high-income students than to that of the low-income group. Ideally, we could test this conjecture about supply-side effects by including demand-shift variables in a multiequation analysis; this is, however, beyond the scope of the present study.

¹⁴This coefficient is computed from information in Manski and Wise's (1983) tables 7.2 and 7.4.

TABLE 3—PRIVATE INSTITUTIONS; DEPENDENT VARIABLE = ENROLLMENT RATE

Variable	Parameter estimate	SE	<i>t</i> for H_0 : parameter = 0
Intercept	0.165	0.019	8.47**
NETCSTPR	-0.036	0.006	-6.27**
TIME ($\times 10^{-3}$)	0.487	0.551	0.88
FEMALE	0.016	0.004	3.80**
MED	-0.028	0.027	-1.02
HIGH	-0.069	0.054	-1.29
NTCSTHPR	0.052	0.012	4.23**
NTCSTMPR	0.023	0.008	2.95**
TIMEHI ($\times 10^{-3}$)	-3.880	2.022	-1.92 ^a
TIMEMED ($\times 10^{-3}$)	0.156	0.802	0.20
FEMHI	0.012	0.009	1.28
FEMMED	0.005	0.005	0.99
Test			χ^2
NETCSTPR + NTCSTHPR = 0			2.22
NETCSTPR + NTCSTMPR = 0			6.65**
TIME + TIMEHI = 0			3.04 ^a
TIME + TIMEMED = 0			1.22
FEMALE + FEMHI = 0			11.46**
FEMALE + FEMMED = 0			45.11**

Notes: $N = 60$; mean enrollment rate = 0.111; root MSE = 0.014; CV = 12.4; $R^2 = 0.94$; adjusted $R^2 = 0.93$.

^aSignificant at $P < 0.10$ level; **significant at $P < 0.01$ level.

TABLE 4—PUBLIC INSTITUTIONS; DEPENDENT VARIABLE = ENROLLMENT RATE

Variable	Parameter estimate	SE	<i>t</i> for H_0 : parameter = 0
Intercept	0.327	0.059	5.50**
NETCSTPU	-0.038	0.034	-1.12
TIME ($\times 10^{-3}$)	-3.646	1.960	-1.86 ^a
FEMALE	0.029	0.009	3.19**
MED	-0.179	0.072	-2.47*
HIGH	-0.256	0.076	-3.37**
NTCSTHPU	0.149	0.038	3.91**
NTCSTMPU	0.098	0.038	2.59*
TIMEHI ($\times 10^{-3}$)	3.209	2.350	1.37
TIMEMED ($\times 10^{-3}$)	2.631	2.246	1.17
FEMHI	-0.007	0.013	-0.58
FEMMED	0.001	0.011	0.12
Test			χ^2
NETCSTPU + NTCSTHPU = 0			43.17**
NETCSTPU + NTCSTMPU = 0			13.84**
TIME + TIMEHI = 0			0.11
TIME + TIMEMED = 0			0.86
FEMALE + FEMHI = 0			5.73*
FEMALE + FEMMED = 0			21.63**

Notes: $N = 60$; mean enrollment rate = 0.318; root MSE = 0.021; CV = 6.56; $R^2 = 0.93$; adjusted $R^2 = 0.91$.

^aSignificant at $P < 0.10$ level; *significant at $P < 0.05$ level; **significant at $P < 0.01$ level.

will be reflected in higher average net cost. Distinguishing between sectors does not completely eliminate this problem, since there is price variation within each sector, but it reduces the problem substantially. The structure of the equations is similar to that in Table 2, which combines public and private enrollment, except that the net-cost variables (NETCSTPU and NETCSTPR, respectively) and the net-cost \times income interaction terms (NTCSTMPU and NTCSTHPU for public middle- and high-incomes and NTCSTMPR and NTCSTHPR for private middle- and high-incomes) are specific to the sector whose enrollment is being explained. It would be natural to test for the significance of variables measuring cross-price effects. Unfortunately, a high correlation between the time series for public and private net costs (on the order of 90 percent) makes it impossible to include both variables in the same equation.

As in the combined equation, all the coefficients in the private and public equations that are significant have the expected sign.¹⁵ For private enrollment, we estimate that a \$100 increase in net cost lowers enrollment by about 6.0 percent for low-income students. In the private-institution equation, the net-cost \times middle-income interaction is positive and significant, implying that the price responsiveness of students from middle-income families is significantly lower than that of students from low-income families. The overall net effect of cost on private enrollment for middle-income families is negative and significant, indicating that, as for low-income students, rises in net cost reduce enrollment for middle-income students. The net-cost \times income interaction

variable for students from high-income families is also positive and significant, indicating that they are less responsive to price. However, the overall net effect of cost increases on high-income private enrollment is not significantly different from zero.

Continuing with the results for private enrollment in Table 3, we find that low-income women have a significantly higher enrollment propensity than low-income men. Moreover, chi-square values indicate that enrollment propensities in private colleges are also significantly higher for middle-income and high-income women than for men of the same income class. We find a 0.34-percentage-point negative and significant time trend for high-income students. The time trends for the low- and middle-income groups are not statistically significant.

Turning to the results for public enrollment in Table 4, we find that the coefficient on net cost for low-income students has the expected negative sign but is not statistically significant at conventional levels. (The point estimate would imply that a \$100 increase in net cost reduces enrollment at public institutions by about 1.6 percent for low-income students.) As expected, the coefficients on the net-cost \times income interactions are both positive and significant. For both middle- and high-income groups, chi-square values indicate that the net effect of cost on enrollment is positive and statistically significant. Again, the FEMALE variable was positive and statistically significant for each income group. The only significant time trend is a small negative one (-0.36 percentage points per year) for low-income students.¹⁶

¹⁵The relative size of the coefficients in the public and private equations may also be of interest. Even taking student aid into account, low-income students on average face higher prices in private institutions. If this implies that private higher education is viewed as a "luxury" by low-income families, then low-income students might be expected to be more sensitive to changes in the price at private than at public institutions. However, we do not find a statistically significant difference between the net-cost coefficients for the two sectors in our sample.

¹⁶In a further refinement of the analysis, we break down net cost into its two components: the published tuition (or sticker price) and the subsidy value of aid (AID). (These results are discussed in detail in McPherson and Schapiro [1990]). This step serves the purposes, first, of shedding light on the relative magnitudes of the aid and sticker-price effects and, second, of pushing the data to see if anomalies or inconsistencies surface. When public and private institutions are combined, the two variables have the expected sign: a higher sticker price lowers enrollment, and more aid raises enrollment. The sticker-price coefficient is statistically significant, but the AID coefficient is not significant at the 10-percent level. As for the magnitudes,

III. Summary and Conclusions

Our most important and reliable finding is that increases in the net cost of attendance have a negative and statistically significant effect on enrollment for white students from low-income families. Moreover, the magnitude of this net-cost effect is reasonable in light of that found in earlier econometric studies of enrollment demand. These results hold for a combined sample of public and private institutions as well as for a subsample limited to private institutions.¹⁷ (For the public sample, the sign was as expected but not significant.) It is not possible to use our data set to test for net-cost effects for blacks or other racial-ethnic groups because of excessive sampling variation in the estimated enrollment rates for these students.

Our finding that the time-series and cross-sectional results for low-income white students are consistent is an important first step in resolving a long-standing controversy in the literature.¹⁸ These results derive from

the absolute values of the two coefficients are not statistically different from each other. When public and private enrollment are considered separately, we find that for low-income students at private institutions, AID has the expected positive sign and is statistically significant; sticker price has the expected negative sign, and is also statistically significant; and again, the coefficients do not differ significantly from each other. For low-income students at public institutions, the two signs are as expected but are not statistically significant.

¹⁷When net cost is broken down into its two components, tuition and subsidy value of aid, the results generally continue to support the finding that the enrollment decisions of low-income students are sensitive to the costs they face.

¹⁸A referee raised the interesting point that the post-1980 behavior of several important variables influencing enrollment differed substantially from their behavior before 1981. In particular, tuition rose quite rapidly after 1980, and the growth of federal student aid slowed substantially. Might the difference between our results and those of Hansen (1983) simply result from our inclusion of post-1980 data that did not exist at the time of Hansen's study? We examined this possibility by estimating our equations for the 1974–1980 time period. The results were essentially quite similar to those for the full period, although some of the coefficients were less precisely estimated. An attempt to obtain estimates for the 1981–1984 period was unsuccessful, an outcome we attribute to the severe limitation imposed by the degrees of freedom.

the fact that we have systematically related changes in net cost to changes in enrollment and have not simply looked at enrollment levels at two points in time. It is important to appreciate that these findings for low-income students would be obscured in an analysis that aggregated over income groups, since our evidence suggests (in line with the findings of cross-sectional studies) that the behavior of these income groups is quite different.

We found a very different picture when we looked at the behavior of more-affluent students. We found no evidence in these data that increases in net cost inhibited enrollment in these income groups. In fact, for the upper-income group, there was a fairly consistent positive effect of net cost on enrollment, which may be interpreted as indicating a tendency for high enrollment demand among affluent students to lead to higher net costs for those students. For middle-income students, we found that net cost did not have a consistent effect on enrollment in our equations.

The above analysis indicates that changes in the net price facing lower-income students have significant effects on their enrollment behavior. An important policy issue, however, is whether changes in federal aid in fact wind up changing net cost. If, for example, increases in federal aid led to decreases in the amount of aid awarded by institutions or to increases in tuition, the effect of aid on net cost would be muted. This issue deserves more systematic treatment than we can give it here. However, findings from a study of the effects of student aid on institutions (McPherson et al., 1989; McPherson and Schapiro, 1991) suggest that these potential offsetting effects may not be empirically important. The time-series evidence on net cost further suggests that periods when federal aid is generous coincide with periods when the net cost facing low-income students is lower. This supports the view that these potential offsets are not important factors.

In sum, a more careful analysis of the historical data has raised serious doubts about the hypothesis that federal student aid has failed to affect enrollment patterns in U.S. higher education significantly over

the past two decades. Our assessment indicates that time-series evidence on the enrollment behavior of low-income white students is quite consistent with the many cross-sectional estimates of aid effects in the literature. While further analysis seems warranted, it is nonetheless clear that policymakers must carefully consider potential enrollment effects when determining aid policy.

REFERENCES

- Bosworth, Barry, Carron, Andrew S. and Rhyne, Elisabeth, *The Economics of Federal Credit Programs*, Washington, DC: Brookings Institution, 1987.
- Campbell, Robert and Siegel, Barry N., "The Demand for Higher Education in the United States, 1919-1964," *American Economic Review*, June 1967, 57, 482-94.
- Gillespie, Donald A. and Carlson, Nancy, *Trends in Student Aid: 1963-83*, Washington, DC: College Board, 1983.
- Hansen, W. Lee, "Impact of Student Financial Aid on Access," in Joseph Froomkin, ed., *The Crisis in Higher Education*, New York: Academy of Political Science, 1983, 84-96.
- Hauptman, Arthur, *Federal Costs for Student Loans: Is There a Role for Institution-Based Lending?*, Washington, DC: American Council on Education, 1985.
- Hight, John E., "The Demand for Higher Education in the United States 1927-72; the Public and Private Institutions," *Journal of Human Resources*, Fall 1975, 10, 512-20.
- Hoenack, Stephen A. and Weiler, William, "Cost-Related Tuition Policies and University Enrollments," *Journal of Human Resources*, Summer 1975, 10, 332-60.
- Katz, Lawrence F. and Murphy, Kevin M., "Changes in Relative Wages, 1963-1987: Supply and Demand Factors," unpublished manuscript, NBER (Cambridge, MA), 1990.
- Leslie, Larry L. and Brinkman, Paul T., "Student Price Response in Higher Education: The Student Demand Studies," *Journal of Higher Education*, March/April 1987, 58, 181-204.
- _____, and _____, *The Economic Value of Higher Education*, New York: Macmillan, 1988.
- Lewis, Gwendolyn L., *Trends in Student Aid, 1980-1988*, Washington, DC: College Board, 1988.
- Manski, Charles F. and Wise, David A., *College Choice in America*, Cambridge, MA: Harvard University Press, 1983.
- McPherson, Michael S., "The Demand for Higher Education," in David Breneman and Chester Finn, eds., *Public Policy and Private Higher Education*, Washington, DC: Brookings Institution, 1978, 143-96.
- _____, *How Can We Tell If Federal Student Aid is Working?*, Washington, DC: College Board, 1988.
- _____, and Schapiro, Morton Owen, "Measuring the Effects of Federal Student Aid: An Assessment of Some Methodological and Empirical Problems," Williams Project on the Economics of Higher Education Discussion Paper No. 4, 1990.
- _____, and _____, *Keeping College Affordable: Government and Educational Opportunity*, Washington, DC: Brookings Institution, 1991 (forthcoming).
- _____, _____, and Winston, Gordon C., "Recent Trends in U.S. Higher Education Costs and Prices: The Role of Government Funding," *American Economic Review*, May 1989 (*Papers and Proceedings*), 79, 253-7.
- White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, 48, 817-38.
- Zemsky, Robert, "The Goal of Diversity: Access and Choice in Academia," mimeo, University of Pennsylvania, 1988.
- American Freshman Survey*, Los Angeles: Cooperative Institutional Research Program, Higher Education Research Institute, Graduate School of Education, University of California, Los Angeles.
- Current Population Survey*, Washington, DC: U.S. Bureau of the Census, U.S. Department of Commerce.

Are Workers Permanently Scarred by Job Displacements?

By CHRISTOPHER J. RUHM*

This paper investigates whether workers suffer lasting "scars" following job displacements. Using David T. Ellwood's (1982) terminology, "scars" represent persistent effects, whereas "blemishes" are transitory adjustments which dissipate over time. More precisely, dislocated individuals are defined as scarred if they continue to earn less or to be unemployed more than their nondisplaced counterparts, even after the conclusion of a several-year adjustment period.

Displacements may have a transitory impact if workers initially obtain unstable positions but later move to more secure jobs or if low wages are received during short-lasting probationary or training periods. These temporary effects, although important, cause less concern than persistent scarring. It is therefore somewhat surprising that, despite the proliferation of recent research studying the consequences of permanent layoffs, relatively little is known about the duration of the associated adjustment problems.¹

Data for this study were obtained for heads of households from the 1969–1982 waves (survey reporting years) of the *Michigan Panel Study of Income Dynamics* (PSID). Displacement status was ascertained for the five base years 1971–1975, with respondents defined as permanently displaced if they terminated jobs as the result of plant closings or layoffs (excluding departures from temporary or seasonal jobs) and failed to return to the original employer by the end of the second full calendar year

following the layoff. Data were collected on unemployment and weekly wages for the three years preceding and four years following the base period. The base year is hereafter referred to as time t , the following four years are referred to as periods $t+1$ through $t+4$, and numbers subtracted from t refer to periods preceding the base year. Information was also assembled indicating whether respondents were displaced in year $t+5$.

The sample includes household heads between the ages of 21 and 65 at time t who participated in the labor force during some part of each of years t through $t+5$ and received positive earnings at some time during period $t-2$ or $t-3$. The analysis is therefore restricted to individuals with fairly strong attachments to the labor force.² To reduce the number of multiple observations for given individuals, data were used only for base years in which the individual was displaced and for a maximum of one randomly chosen base year in which no displacement occurred.³ The sample contains 3,813 person-year observations, 800 for workers losing jobs at time 0 and 3,013 for individuals not displaced during that year.

²The PSID oversamples low-income individuals and has information on relatively few permanent job changers (even with pooling). This raises concern that the results of this study might differ from those obtained using more-representative data. For this reason, regression estimates of postdisplacement unemployment on a commonly defined set of control variables were compared for the PSID and for household heads in the *Displaced Workers Survey*, a supplement to the January 1984 *Current Population Survey*. Coefficients from the two sets of regressions were similar, suggesting that the findings of this paper are reasonably robust.

³The precise selection criteria were as follows: a random variable V was created which, for each individual, had an equal probability of taking integer values $1 \leq V \leq 5$; the base year was then included in the sample if: 1) $\text{YEAR} - 1970 - V = 0$; or 2) if a permanent displacement occurred either at t or $t+5$.

*Assistant Professor, Department of Economics, Boston University. I thank Peter Doeringer, Daniel Hamermesh, and Lori Kletzer for helpful comments on earlier versions of this paper. Funding was received from the U.S. Department of Labor, Bureau of International Labor Affairs. The usual disclaimer applies.

¹See the survey article by Daniel S. Hamermesh (1989) and the edited volume by John T. Addison (1990) for examples of recent work.

I. Estimation Technique

The effect of displacements occurring at time t on employment conditions at $t+n$ can be estimated from regressions of

$$(1) \quad Y_{i,t+n} = \mathbf{X}_{i,t}\alpha + D_{i,t}\beta + \mu_{i,t+n}$$

where i is an individual subscript, Y is the dependent variable (either weeks of unemployment or the natural log of weekly wages), \mathbf{X} is a vector of observable characteristics, D is a dummy variable indicating whether a displacement occurs at time t , and μ is the regression disturbance term. These estimates will be biased, however, if the vector of independent variables fails to control fully for all types of heterogeneity that jointly influence the dependent variable and the probability of displacement. For this reason, two additional specifications were estimated.

In the first, predisplacement wages and predisplacement unemployment ($Y_{i,t-m}$) were included to control for heterogeneity not captured by \mathbf{X} , yielding the regression equation

$$(2) \quad Y_{i,t+n} = \mathbf{X}_{i,t}\alpha + D_{i,t}\beta + Y_{i,t-m}\gamma + \mu_{i,t+n}$$

The second specification included persons involuntarily terminating jobs in year 5 (after the end of the observation period) as a comparison group. These equations were of the form

$$(3) \quad Y_{i,t+n} = \mathbf{X}_{i,t}\alpha + D_{i,t}\beta + D_{i,t+5}\delta + \mu_{i,t+n}$$

with the bias introduced by unobserved heterogeneity captured by the $\hat{\delta}$. The net displacement effect was then calculated as $\hat{\beta} - \hat{\delta}$, where $\hat{\delta}$ is the average value of $\hat{\delta}$ for periods t through $t+2$.⁴ The earnings re-

gressions were estimated using ordinary least squares, and the unemployment equations were estimated as TOBIT models.

There is some ambiguity as to the appropriate choice of lagged dependent variables to be used in equation (2). If a period shortly before the involuntary termination is chosen (e.g., Y_{t-1}) and employment conditions begin to deteriorate prior to permanent separations (as firms institute temporary layoffs or obtain wage concessions), the coefficient on the lagged dependent variable will include a portion of the displacement effect, and $\hat{\beta}$ will understate the impact of permanent layoffs. Conversely, the use of a longer lag period will cause $\hat{\beta}$ to overstate the actual displacement effect if firms strategically terminate workers whose performance has been deteriorating or who previously received high pay, relative to their productivity. This occurs because the displacement coefficient captures the effects of changes that would have been expected even in the absence of mobility.⁵

Preliminary analysis of the comparison group of workers displaced in year $t+5$ revealed that wage and employment adjustments begin two calendar years before permanent layoffs. For this reason, equation (2) was estimated with the dependent variable alternatively lagged one and three periods prior to the base year. Regressions with Y_{t-1} (Y_{t-3}) included are likely to provide lower (upper) bounds on the actual displacement effect.

II. Wage Scars and Unemployment Blemishes

Permanent layoffs are associated with substantially elevated initial unemployment. Between years t and $t+2$, dislocated workers were three times more likely than their counterparts to experience some unemployment (83.2 percent vs. 26.3 percent), aver-

⁴As discussed in the next paragraph, $\hat{\delta}$ will include a portion of the displacement effect in later years, if adjustments begin prior to the job termination. Jacob Mincer (1986) and I (Ruhm, 1990) have previously used subsequent job losers as a comparison group to account for unobserved differences when analyzing the effects of job mobility

⁵Closely analogous concerns are raised in the econometric analysis of training programs, where reductions in employment and earnings are typically observed shortly prior to the beginning of the training period. See Orley Ashenfelter (1978), Ashenfelter and David Card (1985), and Card and Daniel Sullivan (1988) for discussions of these issues.

TABLE 1—POSTDISPLACEMENT WAGE AND UNEMPLOYMENT DIFFERENTIALS

Time period	Unemployment				Weekly wages			
	a	b	c	d	a	b	c	d
t	8.35 (27.28)	8.43 (27.22)	7.10 (25.94)	8.73 (26.99)	-0.1058 (5.39)	-0.1054 (6.49)	-0.0645 (4.64)	-0.1149 (5.73)
$t + 1$	4.32 (17.27)	4.18 (17.15)	3.69 (15.53)	4.64 (19.98)	-0.1751 (8.73)	-0.1664 (9.89)	-0.1365 (9.09)	-0.1820 (8.88)
$t + 2$	2.08 (9.45)	2.11 (9.21)	1.79 (8.29)	2.24 (9.65)	-0.1623 (7.77)	-0.1585 (8.83)	-0.1188 (7.25)	-0.1690 (7.91)
$t + 3$	1.45 (7.00)	1.40 (6.83)	1.16 (5.84)	1.77 (8.07)	-0.1486 (7.67)	-0.1435 (8.03)	-0.1067 (6.43)	-0.1627 (7.17)
$t + 4$	1.27 (5.81)	1.12 (5.62)	0.88 (4.60)	1.85 (8.56)	-0.1473 (6.84)	-0.1402 (7.34)	-0.1117 (6.33)	-0.1698 (7.73)
Control for heterogeneity:	none	Y_{t-3}	Y_{t-1}	D_{t+5}	none	Y_{t-3}	Y_{t-1}	D_{t+5}
				0.86 (2.61)				-0.0434 (1.77)

Notes: See text for explanation of columns a–d. Unemployment equations are estimated using maximum-likelihood TOBIT methods; the wage equations are estimated by ordinary least squares. The unemployment coefficients show the impact of marginal changes in the regressors on actual joblessness and are obtained by evaluating the TOBIT regressions at the independent variable means. The wage coefficients show the impact of base-year displacements on the natural log of weekly wages. Numbers in parentheses are t statistics. Regressions include controls for experience, education, marital status, race, sex, city size, tenure, the survey year, industry, occupation, and age (> 55 years old). Coefficient on D_{t+5} is the average obtained from regressions for years t through $t + 2$.

aged six times as many weeks out of work (23.9 vs. 3.8 weeks), were eight times more probable to be unemployed for more than six months (35.8 percent vs. 4.5 percent), and were jobless for more than one year 16 times as frequently (12.9 percent vs. 0.8 percent).

Dislocated workers were also more than twice as likely to experience wage reductions exceeding 25 percent between $t - 2$ and $t + 2$ (28.6 percent vs. 14.1 percent), and lost more than 10 percent of previous wages 1.6 times as often (40.1 percent vs. 25.0 percent). Although displaced individuals averaged only a 1.5-percent earnings reduction over the four years, they missed out entirely on the 8.4-percent real wage gain obtained by the control group over the same period.

I now turn to considering whether displacements leave lasting scars. Table 1 displays regression results of unemployment

and wage equations estimated for periods t through $t + 4$. For each dependent variable, coefficients in column a are obtained from estimating equation (1), which includes no controls for unobserved heterogeneity. Coefficient in columns b and c are from equation (2) with Y_{t-3} and Y_{t-1} , respectively, incorporated.⁶ Estimates in column d are from equation (3), with the final entry in the column indicating δ . In this case, the effects of base-year layoffs are calculated by subtracting δ from the estimated coefficients on D_t . Displacement impacts in the early years (t through $t + 2$) indicate adjustment costs in the periods immediately following permanent job loss. The major focus of this paper, however, is on longer-run effects, as

⁶If data for period $t - 3$ were unavailable (in Table 1, column b), corresponding information for period $t - 2$ was used.

measured by the coefficients for years $t+3$ and $t+4$.

Permanent layoffs lead to substantial temporary unemployment blemishes but far more enduring earnings scars. Displaced workers were out of work eight weeks more than their observably similar counterparts in the year of the separation, four additional weeks in period $t+1$, and two extra weeks at $t+2$. By year $t+3$ they were jobless only 1.5 weeks more than the peer group, and the $t+4$ increase was just six days (see Table 1, column a). Thus, at least 85 percent of the initial rise in joblessness dissipated prior to year $t+4$.⁷

In contrast, wage effects are large and lasting. The estimates in column a of Table 1 imply that base-year weekly earnings of job losers were 10.0 percent below those of their nondisplaced peers. This understates the reduction caused by dislocations, however, since period t wages partially reflect the (typically higher) pay received on the pre-separation job.⁸ It is therefore not surprising that the period $t+1$ wage differential was 1.6 times greater (16.1 percent) than the base-year effect. More significantly, almost none of the $t+1$ wage reduction dissipated with time. The earnings gap remained at 13.8 percent and 13.7 percent, respectively, in years $t+3$ and $t+4$, which was 85 percent of the disparity predicted for year $t+1$.⁹

The inclusion of controls for unobserved heterogeneity only slightly reduces the measured impact of base-year terminations. As expected, the estimated displacement effect

was somewhat larger when Y_{t-3} was added to the regressions (Table 1, column b) than when Y_{t-1} was included (column c). Even in the latter case, however, unemployment increased temporarily (by 7.1 weeks in year t and by 3.7 weeks in year $t+1$), and the wage loss remained above 10 percent in periods $t+2$ through $t+4$. If anything, addition of the lagged dependent variables strengthens the earlier finding of unemployment blemishes and permanent wage scars. This is seen by noting, that although the unemployment effect fell to around one week by period 4, the upper (lower) bound on the corresponding wage loss remained at 13.1 (10.6) percent.

The coefficient on D_{t+5} , in column d of Table 1, indicates that unobserved heterogeneity accounted for a 0.9-week disparity in annual unemployment and a 4-percent wage differential between displaced and nondisplaced workers. Subtracting this from the coefficients on D_t , the net increase in unemployment was 7.9 weeks at time 0, 3.8 weeks in period $t+1$, but less than 1 week annually in periods $t+3$ and $t+4$. Conversely, the estimated wage loss exceeded 11 percent in each of years $t+2$ through $t+4$, which was almost 90 percent as large as the maximum reduction observed in period $t+1$. Interestingly, these displacement effects are quite close to the lower-bound estimates presented in column c of Table 1 and again indicate transitory employment shocks but lasting earnings changes.

III. Summary

Although permanent loss of jobs leads to unemployment distinguished by its long duration, there is no evidence that these initial difficulties translate into lasting scars. Conversely, involuntary terminations typically result in a significant loss of long-term earnings potential. Four years after displacement, job losers are out of work only one week more than their nondisplaced counterparts but continue to earn 10–13 percent less. The lasting wage reductions suggest significant worker attachments to specific jobs. Future research is needed to investigate the sources of these attachments.

⁷Ellwood (1982) also finds that current unemployment has only a small effect on future joblessness. His results are for teenagers.

⁸For example, individuals displaced in July worked more than half the base year in the predisplacement position.

⁹Evidence of persistent wage losses has also recently been obtained by Robert Topel (1989). The wage-equation model in column a of Table 1 was also estimated using the standard two-stage correction for sample-selection bias. Except for period t , for which the estimates indicated that workers with low potential wages were more likely to be reemployed, the ordinary and two-stage least-squares estimates were virtually identical. The latter, if anything, showed greater persistence of wage losses.

APPENDIX

Data Set and Variable Construction

Data were taken from the 1969–1982 waves of the *Panel Study of Income Dynamics* (PSID). The PSID was obtained from the Inter-University Consortium for Political and Social Research (ICPSR); requests for descriptions of the data set and for the raw data should be directed to the ICPSR.

Analysis was restricted to household heads between the ages of 21 and 65 during the 1971–1975 base years who: 1) participated in the labor force during some part of the base year (t) and the following five years; 2) received earnings during at least one of the years $t-3$ or $t-2$; and 3) fulfilled the sampling criteria specified in footnote 3. Respondents were excluded if the head of household had changed during the relevant sample period or if they retired during the five years subsequent to the base year.

All data manipulations and analyses were performed using SAS (SAS Institute, 1985) on the Boston University mainframe computer, with the exception of the TOBIT regressions which were estimated using LIMDEP (William H. Green, 1986).

Individuals were defined to be displaced if they involuntarily terminated jobs in the base period [or in year 5 for the comparison group described in equation (3)] and failed to return to the old employer by the end of the second full calendar year following the separation. Dependent variables used in the analysis were weeks of calendar-year unemployment and the log of weekly wages. The latter were calculated by taking the natural logarithm of real annual earnings divided by number of weeks worked, with annual earnings adjusted to 1972 prices using the GNP deflator. Missing values were assigned for periods in which the natural log of weekly wage was less than 3 (this corresponds to weekly wages of less than \$20).

Except for the lagged dependent variables, the regressors are all dummy variables and refer to individual or job characteristics in the base year. They equal 1 if the following are true (and 0 otherwise).

Experience: Labor market experience is less than 10 years. Data on actual labor market experience is available and used beginning in 1974. For earlier years, experience is calculated as 1974 experience less the difference between 1974 and the survey date (e.g., in 1972, experience was set equal to 1974 experience less 2 years).

School: Education is less than or equal to 12 years.

Married: Respondent is married or permanently cohabiting.

Female: Respondent is female.

Nonwhite: Respondent is black or Spanish-American.

Age: Age is greater than 55 years old.

City1: City size is greater than 100,000.

City2: City size is less than 25,000.

Blue Collar: One-digit predisplacement occupation code is 5, 6, or 7.

Manufacturing: One-digit predisplacement industry code is 3 or 4.

Professional: One-digit predisplacement occupation code is 1 or 2.

Tenure1: Predisplacement seniority is 1–3 years.

Tenure2: Predisplacement seniority is 4–9 years.

Tenure3: Predisplacement seniority is 10–19 years.

Tenure4: Predisplacement seniority is at least 20 years.

Survey Year: Four dummy variables were included indicating the survey years 1972–1975 (1971 is the excluded category.)

REFERENCES

- Addison, John T., *Job Displacement: Consequences and Implications for Policy*, Detroit: Wayne State University Press, 1990 forthcoming.
- Ashenfelter, Orley, "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, February 1978, 60, 47–57.
- _____ and Card, David, "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, November 1985, 67, 648–60.

- Card, David and Sullivan, Daniel, "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Unemployment," *Econometrica*, May 1988, 56, 497-530.
- Ellwood, David T., "Teenage Unemployment: Permanent Scars or Temporary Blemishes?" in R. B. Freeman and D. A. Wise, eds., *The Youth Labor Market: Its Nature, Causes and Consequences*, Chicago: University of Chicago Press, 1982, 349-85.
- Green, William H., *LIMDEP*, Version 4, New York: Econometric Software, 1986.
- Hamermesh, Daniel S., "What Do We Know About Worker Displacement in the U.S.," *Industrial Relations*, Winter 1989, 28, 51-9.
- Mincer, Jacob, "Wage Changes in Job Changes," in Ronald Ehrenberg, ed., *Research in Labor Economics*, Vol. 8, Greenwich, CT: JAI Press, 1986, 171-97.
- Ruhm, Christopher J., "The Time Profile of Post-Displacement Earnings and Unemployment Changes," in John T. Addison, ed., *Job Displacement: Consequences and Implications for Policy*, Detroit: Wayne State University Press, 1990 forthcoming.
- SAS Institute, *User's Guide: Statistics*, Version 5 Ed., Cary, NC: SAS Inst., Inc., 1985.
- Topel, Robert, "Specific Capital and Unemployment: Measuring the Costs and Consequences of Job Loss," mimeo, University of Chicago, 1989.
- Displaced Workers Survey*, Supplement to *Current Population Survey*, Washington, DC: U.S. Bureau of the Census, U.S. Department of Commerce, January, 1984.
- Panel Study of Income Dynamics: Procedures and Tape Codes*, Ann Arbor, MI: Institute for Social Research, University of Michigan, annual 1969-1982.

Structural Determinants of Real Exchange Rates and National Price Levels: Some Empirical Evidence

By JEFFREY H. BERGSTRAND*

Contrary to the long-held notion of purchasing power parity (PPP), economists have found systematic evidence that the general level of prices across countries at a point in time varies dramatically. Irving B. Kravis, Alan W. Heston, and Robert Summers (1982), for example, report that some countries' national price levels are no more than *one-third* the U.S. price level. Extensions of this work show that such departures from PPP have persisted for decades.

Recently, efforts have been made to explain systematically these persistent, or structural, departures from PPP. Pioneering work by Kravis and Robert E. Lipsey (1983, 1987, 1988) has demonstrated that a positive correlation between the price level and (real) per capita gross domestic product is robust across numerous cross-sectional specifications. For instance, using data from Kravis et al. (1982 table 6-12), 87 percent of the variation in national price levels (PL) of 21 countries¹ in 1975 is explained by per capita GDP (y) and a constant:

$$(1) \quad \ln(PL) = 2.20 + 0.56 \ln y$$

(12.25) (11.72)

($\bar{R}^2 = 0.87$, RMSE = 0.18; t statistics in parentheses).

*Department of Finance and Business Economics, College of Business Administration, University of Notre Dame, Notre Dame, IN 46556. I am grateful to Ron Balvers, Tom Bundt, Jim Hartigan, Ron Jones, Norman Miller, J. David Richardson, Rich Sheehan, and two anonymous referees for helpful comments on an earlier draft, which was presented at the Western Economics Association annual meeting, June 1989, in Lake Tahoe, Nevada. Financial support from the Center for Research in Business at the University of Notre Dame, with funding from the Transamerica Fund, is gratefully acknowledged.

¹The 21 countries are India, Sri Lanka, Thailand, the Philippines, Korea, Columbia, Jamaica, Brazil, Yugoslavia, Ireland, Italy, Spain, the United Kingdom,

Although Christopher Clague (1986) showed that other "structural" characteristics (such as the trade balance, tourism receipts' share of GDP, and minerals' share of GDP) have significant explanatory power when also included, *why* per capita GDP has such a robust empirical correlation to the price level and *what* economic factor(s) it represents have not yet been determined. In the study of structural determinants of the price level, the two predominant competing theories for per capita GDP's role are the productivity-differentials (or Ricardian) model usually attributed to Bela Balassa (1964) and Paul A. Samuelson (1964) and the relative-factor-endowments (or Heckscher-Ohlin) model discussed in Jagdish N. Bhagwati (1984).

The purpose of this study is to distinguish empirically these two competing supply-oriented hypotheses, along with a possible third *demand*-oriented hypothesis that has received virtually no attention. This third hypothesis suggests that, assuming nonhomothetic tastes,² price levels are higher in countries with higher per capita GDP's because nontraded services are luxuries in consumption while traded commodities are necessities. The empirical evidence in the second section suggests that the null hypothesis that per capita income has no effect on the national price level via this demand channel is rejected.

Japan, Austria, the Netherlands, Belgium, France, Denmark, Germany, and the United States. These countries were selected for constraints that will become apparent in Section II.

²Nonhomothetic tastes imply that the income-expansion path through the indifference curves of the representative consumer is not a straight line through the origin, generating an income elasticity of demand greater (less) than 1 for the nontraded service (traded commodity).

I. Theoretical Issues

Across countries, price levels are expected to be positively associated with [real per capita] income because prices of nontradeables [mainly services] are higher relative to prices of tradeables [mainly commodities] in rich countries than in poor countries.

[Kravis and Lipsey, 1988 p. 474]³

Consider the national price level as decomposable into (nontraded) services' prices and (traded) commodities' prices. According to the productivity-differentials model, rich countries are believed to have absolute productivity advantages in both services and commodities, but a relative productivity advantage in commodities. Consequently, the relative price of services to commodities (henceforth, the "relative price level" or "real exchange rate") will be higher in countries with larger per capita incomes. Since commodity arbitrage will tend to equilibrate commodities' prices across countries, the national price level will tend to be higher in rich countries, since their price of services relative to commodities is higher. One would then expect to find a similar high correlation between the relative price level (p) and per capita GDP across countries. Indeed, using the same data set as for regression equation (1), 85 percent of the variation in relative prices is explained by per capita income and a constant:

$$(2) \ln p = 2.20 + 0.50 \ln y \\ (12.48) \quad (10.58)$$

($\bar{R}^2 = 0.85$, RMSE = 0.17; t statistics in parentheses).

The predominant alternative explanation to productivity differentials involves relative factor endowments. In the two-good, two-

factor, relative-factor-endowments model, services (commodities) are assumed to be relatively labor-intensive (capital-intensive) in production. Relatively capital-abundant rich countries will have a comparative advantage in producing commodities, so that the price of services relative to commodities, and thus the national price level, will be higher in countries with larger per capita income.

However, little attention has been devoted to a Linder-type hypothesis as an alternative demand-side explanation for why per capita GDP is positively correlated with the real exchange rate and national price level. Although Rudiger Dornbusch (1988) and J. Peter Neary (1988) noted that shifts in tastes as well as in technologies and relative factor endowments can change the real exchange rate, no one has attempted to link explicitly differences in tastes across countries with differences in their per capita GDP's (as suggested in Linder [1961]) and differences in their real exchange rates. To illustrate these linkages, consider the following model (see Neary [1988] for a more general discussion using trade-expenditure functions).

A. Demand

Staffan Burenstam Linder (1961 p. 94) suggested that a "whole array of factors influences the demand structure of a country," but that per capita income was likely to be the "most important single factor." He argued:

At higher [real per capita] incomes, products of different kinds, although filling the same basic needs, are likely to replace less sophisticated types of products; furthermore, products filling new needs are added. . . . But the more we divide total production into subgroups, the greater will be the variations in income elasticity. [pp. 94-5]

I formalize the Linder claim that per capita income has a dominant influence on the structure of demand by assuming the following nonhomothetic, nested Cobb-Douglas-Stone-Geary utility function for the

³Bracketed terms added. In the Kravis et al. (1982) data, the empirical distinction between nontradeables and services and between tradeables and commodities is fairly minor and rests entirely upon the treatment of construction. Tradeables consist of all commodities except construction; nontradeables consist of all services plus construction (see Kravis et al., 1982 p. 193). Consequently, reference here will be made to non-traded services and traded commodities.

representative consumer-worker:

$$(3) \quad u = (x_T - \bar{x}_T)^\delta (x_N - \bar{x}_N)^{1-\delta} \\ 0 < \delta < 1$$

where x_T (x_N) is the amount consumed of the traded commodity (nontraded service) and \bar{x}_T (\bar{x}_N) is an exogenous minimum-consumption requirement that exists for the traded commodity (nontraded service), common to the Stone-Geary utility function. Ready examples of traded commodities and nontraded services that would have minimum per capita consumption requirements are food and government-provided police and fire services, respectively. Assume the budget constraint

$$(4) \quad y = x_T + px_N$$

where y is real income of the representative consumer-worker and p is the relative price of the nontraded service, both expressed in terms of the traded commodity (the numeraire).

This utility structure yields nonunitary income elasticities of demand for the two products. Maximizing (3) subject to (4) yields first-order conditions solvable for demand functions:

$$(5) \quad x_N = (1 - \delta)p^{-1}(y - \bar{x}_T) + \delta\bar{x}_N$$

$$(6) \quad x_T = \delta y + (1 - \delta)\bar{x}_T - \delta p\bar{x}_N$$

The differing income elasticities of demand implied by this structure for the two products are made more transparent following some mathematical manipulation to yield

$$(7) \quad \hat{x}_N = - \left(1 - \frac{\delta\bar{x}_N}{x_N} \right) \hat{p} \\ + \left(1 + \frac{(1 - \delta)\bar{x}_T - \delta p\bar{x}_N}{px_N} \right) \hat{y}$$

$$(8) \quad \hat{x}_T = - \left(\frac{\delta p\bar{x}_N}{x_T} \right) \hat{p} \\ + \left(1 - \frac{(1 - \delta)\bar{x}_T - \delta p\bar{x}_N}{x_T} \right) \hat{y}$$

where \hat{x} denotes $dx/x = d(\ln x)$. In the cross-country context of this paper, \hat{x} is interpreted as a percentage difference between two countries. For example, a 1-percent-higher per capita income in country B relative to country A will cause B's per capita demand for the nontraded service to be $1 + [(1 - \delta)\bar{x}_T - \delta p\bar{x}_N]/px_N$ percent higher than A's and will cause B's per capita demand for the traded commodity to be $1 - [(1 - \delta)\bar{x}_T - \delta p\bar{x}_N]/x_T$ percent higher than A's. The nontraded service (traded commodity) will be the luxury (necessity) in consumption if the parameter-weighted minimum-consumption requirement for the traded commodity, $(1 - \delta)\bar{x}_T$, exceeds that for the nontraded service (expressed in the numeraire), $\delta p\bar{x}_N$, and vice versa.

The demand for the nontraded service relative to the traded commodity (X) is

$$(9) \quad \hat{X} = (\hat{X}_N/\hat{X}_T) = (\hat{x}_N/\hat{x}_T) = \hat{x}_N - \hat{x}_T \\ = -\sigma_D \hat{p} \\ + \left(\frac{x_T + px_N}{x_T px_N} [(1 - \delta)\bar{x}_T - \delta p\bar{x}_N] \right) \hat{y}$$

where X_N (X_T) denotes aggregate demand for the nontraded service (traded commodity) and σ_D is the elasticity of substitution in consumption; formally, $\sigma_D = 1 - (\delta\bar{x}_N/x_N) - (\delta p\bar{x}_N/x_T)$, which is likely to be positive and close to 1. Per capita GDP's coefficient may be positive or negative; a 1-percent-higher per capita income in country B will cause B's relative demand for the nontraded service to be higher (lower) than A's if the weighted minimum-consumption requirement for the traded commodity is greater (less) than that for the nontraded service.

Is there reason to believe a priori that the minimum-consumption requirement per capita for traded commodities exceeds that for nontraded services? The theoretical model reveals no such presumption. However, casual observation of per capita consumption patterns of the poorest countries in Kravis et al.'s (1982) data set (group I) suggests that the minimum-consumption re-

quirement for commodities is likely to dwarf that for services. The only product group in Kravis et al. (1982) that included commodities (services) *exclusively* was food (government compensation for services provided). In terms of international prices, group I's per capita GDP in 1975 was only 9 percent of U.S. per capita GDP; yet 38 percent of this group's low per capita GDP was spent on food, while only 5 percent was spent on government compensation for services provided. Also, Linda C. Hunter and James R. Markusen (1988) used the same data set to estimate linear expenditure systems by Kravis et al.'s (1982) product groups; the results suggested that nontraded services (traded commodities) had income elasticities greater (less) than 1. Nevertheless, the empirical results in Section II will systematically reveal whether the coefficient on \hat{y} is positive or negative.

Finally, for the special case in which \bar{x}_N is zero, the coefficient on \hat{y} is positive, and σ_D equals 1. This will be of interest for Section II.

B. Supply

The purpose of this section is to motivate a function for the supply of nontraded services relative to traded commodities in the representative country. I assume a standard simple-general-equilibrium framework similar to that in Ronald W. Jones (1965) for production of these two goods, using two factors: capital (K) and consumer-workers (L), the endowment of which is fixed at a point in time for each country. Perfectly competitive firms are assumed to minimize costs given the constant-returns-to-scale technology, yielding the optimum input requirements per unit of output:

$$(10) \quad \beta = \begin{bmatrix} \beta_{LN} & \beta_{LT} \\ \beta_{KN} & \beta_{KT} \end{bmatrix}.$$

Each β_{ij} ($i = L, K$; $j = N, T$) is a function of the relative factor price (i.e., the wage rate [W] relative to the rental rate on capital [R]) and the state of productivity (τ_{ij}) in the country. An assumption of full employment

of both factors requires

$$(11) \quad \beta_{LN}X_N + \beta_{LT}X_T = L$$

$$(12) \quad \beta_{KN}X_N + \beta_{KT}X_T = K$$

where L (K) is the overall endowment of labor (capital) and X_N (X_T) is aggregate production of the nontraded service (traded commodity). In a competitive equilibrium with both goods produced, unit costs must reflect market prices of the goods:

$$(13) \quad \beta_{LN}W + \beta_{KN}R = P_N$$

$$(14) \quad \beta_{LT}W + \beta_{KT}R = P_T$$

where all factor prices (W, R) and goods prices (P_N, P_T) are expressed in terms of a monetary unit, as in Jones (1965).

The production framework follows closely sections 2, 3, and 9 in Jones (1965). Hence, derivations for the solution need not be reproduced but are available from the author upon request. With some mathematical manipulation, the production framework can be solved for the supply of nontraded services relative to traded commodities (X) as a function of their relative price level (p), the capital:labor endowment ratio ($k = K/L$), and the level of productivity in traded commodities relative to nontraded services (Π):

$$(15) \quad \hat{X} = \sigma_S \hat{p} - (1/|\lambda|) \hat{k} - (1 + \sigma_S) \hat{\Pi}$$

where the level of productivity in each industry is a weighted average of the level of productivity of each factor in that industry (τ_{ij}), $|\lambda| = \beta_{LN}X_N/L - \beta_{KN}X_N/K = \beta_{KT}X_T/K - \beta_{LT}X_T/L$, σ_S is the elasticity of substitution between goods in production (along the transformation schedule) as in Jones (1965), and $\sigma_S > 0$. In the cross-country context of this paper, a 1-percent-higher level of productivity in traded commodities relative to nontraded services in B compared with A will cause B's relative supply of nontraded services to traded commodities to be $1 + \sigma_S$ percent lower than A's. The coefficient for the capital:labor ratio is ambiguously signed, depending upon rela-

tive factor intensities in production. If nontraded services are relatively labor-intensive in production (i.e., $|\lambda| > 0$), a 1-percent-higher capital:labor ratio in B relative to A will cause B's supply of nontraded services relative to traded commodities to be lower than A's.

C. Equilibrium

Demand function (9) and supply function (15) can be solved for the equilibrium relative price level (or real exchange rate), p , and the equilibrium relative output level, X , in the representative country:

$$(16) \quad \hat{p} = \frac{1 + \sigma_s}{\sigma_D + \sigma_s} \hat{\Pi} + \frac{1}{(\sigma_D + \sigma_s)|\lambda|} \hat{k} + \frac{[(1 - \delta)\bar{x}_T - \delta p\bar{x}_N](x_T + px_N)}{(\sigma_D + \sigma_s)(x_T px_N)} \hat{y}$$

$$(17) \quad \hat{X} = -\frac{\sigma_D(1 + \sigma_s)}{\sigma_D + \sigma_s} \hat{\Pi} - \frac{\sigma_D}{(\sigma_D + \sigma_s)|\lambda|} \hat{k} + \frac{\sigma_s[(1 - \delta)\bar{x}_T - \delta p\bar{x}_N](x_T + px_N)}{(\sigma_D + \sigma_s)(x_T px_N)} \hat{y}.$$

Equation (16) demonstrates how the productivity-differentials, relative-factor-endowments, and Linder hypotheses are all potentially relevant for explaining variation across countries in the equilibrium relative price of nontraded services to traded commodities (and the general price level). A 1-percent-higher productivity in traded commodities relative to that in nontraded services in country B compared with country A will cause B's relative price level to be $(1 + \sigma_s)/(\sigma_D + \sigma_s)$ percent higher than A's, supporting the productivity-differentials model. A 1-percent-higher capital:labor ratio in B relative to A will cause B's relative price level to be $1/[(\sigma_D + \sigma_s)|\lambda|]$ -percent higher than A's if nontraded services are relatively labor-intensive in production ($|\lambda| > 0$), supporting the relative-factor-endowments hypothesis. A 1-percent-higher per capita income in B relative to A will cause B's relative price level to be higher if the weighted minimum-consumption require-

ment for traded commodities exceeds that for nontraded services, implying an income elasticity of demand for nontradeables (tradeables) greater (less) than one.

Thus, all three hypotheses potentially can explain structural variation in the real exchange rate. Since per capita income is correlated positively with the capital:labor ratio and the level of productivity in commodities relative to services across countries, earlier studies have not tried to distinguish empirically among the relative importances of these three channels. However, if capital:labor ratios and relative-productivity measures are available, a ready method of distinguishing between the demand and supply roles of per capita income is to examine empirically the cross-country relationship between per capita income and the output of nontraded services relative to traded commodities, as demonstrated by equation (17). A 1-percent-higher productivity in traded commodities relative to nontraded services in B compared with A will cause B's output of nontraded services relative to traded commodities to be $\sigma_D(1 + \sigma_s)/(\sigma_D + \sigma_s)$ percent lower than A's. A 1-percent-higher capital:labor ratio in B relative to A will cause B's output of nontraded services relative to traded commodities to be $\sigma_D/[(\sigma_D + \sigma_s)|\lambda|]$ percent lower than A's. However, a 1-percent-higher per capita income in B relative to A will cause B's relative output of nontraded services to traded commodities to be higher than A's if, to be consistent with equation (16), the income elasticity of demand for nontradeables (tradeables) is greater (less) than one.

The theoretical arguments are illustrated in Figure 1. Country B might have a higher relative price level than A because B has a higher productivity in traded commodities relative to nontraded services or a higher capital:labor ratio (assuming nontraded services are labor intensive) or both, causing a lower supply of nontraded services relative to traded commodities. However, if nontraded services (traded commodities) are luxuries (necessities) in consumption, higher-per-capita-income country B might have a higher relative price level than A

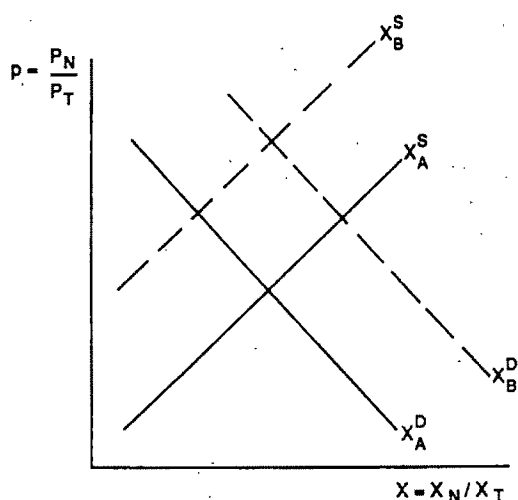


FIGURE 1. RELATIVE DEMAND AND RELATIVE SUPPLY CURVES FOR COUNTRIES A AND B

because of a higher *relative demand* for nontraded services. Only by examining empirically both reduced forms or their underlying structural equations could one hope to disentangle these three influences; these are examined next.

II. Empirical Results

Recent estimates of capital and labor endowments and of levels of productivity in commodities and in services across countries now make it possible to distinguish empirically among the three competing explanations for the robust positive correlation between countries' price levels and their per capita GDP's. Edward E. Leamer (1984) provides measures of capital and labor endowments for numerous countries circa 1975, of which 23 countries overlap with the 34 in Kravis et al. (1982) and Kravis and Lipsey (1983). The technique of Kravis et al. (1983) could be used to approximate the level of productivity in commodities relative to services for 21 of these countries.⁴

⁴This explains the 21 countries for regressions (1) and (2). Capital and labor (LABOR1) for 1975 are from Leamer (1984 appendix table B.1). The level of productivity in (traded) commodities relative to (non-traded) services, $\Pi = \Pi_T / \Pi_N$, is approximated by the

A. Reduced-Form Estimates

Given these 1975 estimates of capital: labor endowment ratios (k) and levels of productivity in commodities relative to services (Π), econometric analogues to reduced-form equations (16) and (17) could be estimated using ordinary least squares (OLS). Estimation of the log-linear version of (16) yields:⁵

$$(18) \quad \ln p = -2.81 + 0.17 \ln \Pi + 0.23 \ln k + 0.18 \ln y$$

(12.68) (2.25) (2.84) (1.67)

($\bar{R}^2 = 0.90$, RMSE = 0.14; t statistics in parentheses). Equation (18) suggests that each supply-oriented hypothesis for the relationship between the real exchange rate and per capita GDP has partial explanatory power. The level of productivity in commodities relative to services has the expected positive effect on the price of services relative to commodities, according to the productivity-differentials hypothesis; the coefficient estimate is statistically significant at the 2.5-percent level (one-tailed t test).⁶

ratio of national output in commodities industries to the level of employment in commodities industries divided by the ratio of national output in services industries to the level of employment in services industries, that is, $(X_T / L_T) / (X_N / L_N)$, the inverse of the calculation in Kravis et al. (1983) for the level of productivity in services relative to commodities. Employment data are from the International Labour Organization's (1979) *Year Book of Labour Statistics*, as in Kravis et al. (1983). The output of commodities relative to services and all other data are from table 6-12 in Kravis et al. (1982).

⁵Treating the coefficients of $\hat{\Pi}$, \hat{k} , and \hat{y} in (16) and (17) as constants, indefinite integration of those equations yields log-linear forms with constants appended.

⁶One referee noted that differences across countries in $(X_T / L_T) / (X_N / L_N)$, the proxy for Π_T / Π_N , could largely reflect differences across countries in their capital per unit of labor in (traded) commodities relative to (nontraded) services—the latter denoted k_T / k_N —rather than differences in Π_T / Π_N . As noted in Section I, each country's β_{LT} ($= L_T / X_T$) and β_{LN} ($= L_N / X_N$) are negative functions of the country's states of productivity (τ_{LT} and τ_{LN} , respectively) and of the country's relative factor price (W / R). Assuming

The capital:labor endowment ratio has a positive effect on the relative price level, suggesting that services are relatively labor-intensive in production, consistent with the relative-factor-endowments hypothesis; the coefficient estimate is significant at the 1-percent level. The presence of both k and Π has significantly eroded the explanatory power of per capita income.⁷

However, real per capita income is still positively related to the price of services relative to commodities, albeit only at the 10-percent significance level. In the context of the model, the remaining statistical significance of this coefficient suggests that per capita income may also be influencing the relative price level through demand. How-

ever, this positive estimate is also consistent with per capita GDP having some residual influence through supply on the real exchange rate unrelated to capital:labor ratios and relative productivity levels. A ready method of distinguishing empirically between a potential demand or supply role for per capita income is to examine the cross-country relationships between the output of services relative to commodities and Π , k , and y , as suggested by equation (17).

Ordinary-least-squares estimation of the log-linear version of (17) yields

$$(19) \quad \ln X = 0.45 - 0.10 \ln \Pi - 0.26 \ln k \\ (1.42) \quad (0.89) \quad (2.17) \\ + 0.26 \ln y \\ (1.73)$$

Hicks neutrality, $\hat{\tau}_{LN} - \hat{\tau}_{LT} = \hat{\Pi}_T - \hat{\Pi}_N = \hat{\Pi}$, implying that $(X_T/L_T)/(X_N/L_N)$ is a positive function of $\hat{\Pi}$; yet $(X_T/L_T)/(X_N/L_N)$ could also be reflecting (k_T/k_N) if countries are using different cost-minimizing relative factor intensities owing to differing wage:rental ratios [variation in which is not already explained by capital:labor endowment ratios, another right-hand-side variable in eq. (18)]. However, this latter possibility does not seem to be empirically confirmed. While calculations of k_T/k_N for all 21 countries was beyond the scope of this study, ready estimates of such intensities were available for six "country groups" (I-VI), reported in Kravis et al. (1983 table 12). Those authors found no systematic pattern for k_T/k_N across country groups. Moreover, I found no systematic correlations of k_T/k_N with respective observations on $(X_T/L_T)/(X_N/L_N)$, y , k , p , or X for these groups.

Note that this specification differs from that in Clague (1986), which retained per capita GDP as a proxy for (non-natural-resource) relative factor abundances and relative productivity levels. Clague notes, "Ideally, we would like to treat real [per capita] income as an endogenous variable, but since we cannot measure resource endowments per capita, efficiency levels, and other determinants of real income, we are forced to use [per capita income] as an exogenous variable" (p. 320). Not surprisingly, a regression of per capita GDP on k and Π reveals strong evidence for interpreting per capita GDP as a proxy for capital:labor ratios and for levels of productivity in commodities relative to services, respectively. The estimated regression for the same 21 countries is

$$\ln y = -0.31 + 0.68 \ln k + 0.31 \ln \Pi \\ (0.63) \quad (7.88) \quad (2.05)$$

($\bar{R}^2 = 0.85$, RMSE = 0.31; t statistics in parentheses).

($\bar{R}^2 = 0.13$, RMSE = 0.20; t statistics in parentheses). Consistent with the relative-factor-endowments and productivity-differentials models as well as equation (18), a higher level of productivity in commodities relative to services or a higher capital:labor ratio is associated with a lower output of services relative to commodities, although the coefficient estimate for Π is not statistically significant. Moreover, a higher per capita income is associated with a higher relative output of services to commodities. In the context of the model, this result suggests that a higher per capita income causes a higher relative price level because of a greater demand for nontraded services relative to traded commodities. The coefficient estimate for y is statistically significant at the 10-percent level (one-tailed t test).

B. Simultaneous-Equations Estimates

Although the results presented in reduced-form equations (18) and (19) are encouraging to the hypothesis that the positive relationship between per capita income and the relative price level is also attributable to a Linder-type demand channel, the statistical evidence is far from conclusive. Moreover, the coefficients from the reduced-form equations are nonlinear combinations of the structural parameters from the underlying

structural equations. Since this particular model is overidentified, the underlying structural parameter estimates for the relative demand and supply variables cannot be determined immediately from reduced-form equations (18) and (19).⁸

However, two-stage least squares (2SLS) conveniently enables determination of the structural parameter estimates of theoretical equations (9) and (15).⁹ Two-stage least-squares estimation of log-linear versions of (9) and (15) yields the following demand (D) and supply (S) equations for the relative output of services to commodities:

$$(20) \quad \ln X^D = -1.96 - 0.90 \ln p$$

(1.87) (2.10)

$$+ 0.40 \ln y$$

(1.79)

$$(21) \quad \ln X^S = 4.61 + 1.48 \ln p$$

(1.15) (1.07)

$$- 0.35 \ln \Pi - 0.60 \ln k$$

(1.00) (1.20)

(RMSE = 0.11 and 0.33, respectively; *t* statistics in parentheses).

Several points are worth noting. First, empirical implementation of theoretical structural equations (9) and (15) yields results consistent with the Linder-type demand hypothesis. The demand for services relative to commodities is significantly (at the 5-percent level, one-tailed *t* test) related to per capita income, holding constant the influences of relative factor endowments and productivity differentials on relative supply. In the context of the model, this suggests that the income elasticity of demand for nontraded services (traded commodities) is greater (less) than one.

⁸This system would be just identified if one of the supply variables were deleted.

⁹The *t* statistics reported are asymptotically valid. Given that there are only three exogenous variables, 2SLS estimation can only be used appropriately here to determine *relative* outputs and prices.

Second, the estimate of the elasticity of substitution in demand (σ_D), 0.9, is statistically significant and has a plausible magnitude. As noted in Section I, for the special case in which the minimum consumption requirement for nontraded services (\bar{x}_N) is zero, σ_D should equal 1 or $-\sigma_D = -1$. In fact, the linear restriction that the coefficient estimate equals -1 could not be rejected even at the 10-percent significance level ($F_{[1, 18]} = 0.056$). This suggests that \bar{x}_N is not significantly different from zero, consistent with earlier observations noted in Section I.

Third, the relative price level, the level of productivity in services relative to commodities, and the capital:labor endowment ratio all have the expected relationships with the supply of services relative to commodities, although their coefficient estimates are not significant. The linear restriction on this supply equation's coefficient estimates implied by theoretical supply function (15)—that the coefficients of the price and Π variables should sum to -1 —was tested also. This linear restriction was rejected at the 10-percent significance level, but could not be rejected at the 5-percent level ($F_{[1, 17]} = 3.838$).

Fourth, as noted in footnote 2, the empirical distinction between services and non-tradeables and between commodities and tradeables is a fairly minor one, resting on the treatment of construction activity. Equations (18)–(21) were also estimated using nontradeables/tradeables data instead of services/commodities data. The results are similar but are omitted here for brevity; these are available from the author upon request.

In concluding this section, I emphasize a key empirical result that has been omitted in this literature: the null hypothesis that per capita income has no effect on real exchange rates and national price levels *via demand* is rejected.

III. Conclusions and Policy Implications

This study has provided empirical evidence that the systematic cross-country relationship between real per capita incomes

and national price levels (or real exchange rates), commonly attributed to the supply-oriented productivity-differentials and relative-factor-endowments hypotheses, can also be attributed partly to a demand-oriented "Linder-type" hypothesis. Assuming nonhomothetic tastes, countries with higher real per capita income will exhibit, in equilibrium, stronger demand for nontraded services relative to traded commodities, raising their relative price. Empirical evidence showed that, even when differences across countries in capital:labor endowment ratios and levels of productivity in commodities relative to services are accounted for, real per capita income still has a significant positive correlation with relative price levels and outputs. The results are consistent with recent studies emphasizing the importance of nonhomotheticity of tastes for explaining consumption and trade patterns across countries (cf. Jerry G. Thursby and Marie C. Thursby, 1987; Hunter and Markusen, 1988). Yet with only 21 usable observations from the Kravis et al. (1982) data set, the results are only encouraging, not conclusive, and more empirical research along these lines seems warranted.

Finally, two policy implications are raised by the issues addressed in this study. First, some of the U.S. dollar's dramatic real appreciation vis-à-vis foreign currencies between 1980 and 1985 and the subsequent large real depreciation from 1985 to 1988 should not be attributed to "disequilibrium," or exchange-rate misalignment. Per capita GDP of most other major industrialized countries fell relative to that of the United States in the first half of the 1980's and rose subsequently, although not nearly by the magnitude of actual real exchange-rate changes. Second, the Plaza Accord of September 1985, which indicated that the major industrialized countries' governments coordinated an effort to facilitate the U.S. dollar's real depreciation toward some "zone," and massive 1988 coordinated central bank interventions, which revealed that the same countries coordinated an effort to prevent the dollar's depreciation below this zone, both suggest that the governments have some notion of the "equilibrium" real

exchange rate. The results here suggest that relative productivity levels, capital:labor ratios and tastes can explain as much as 90 percent of the variation across countries in real exchange rates. Consequently, the variation in such diverse real economic variables over time will tend to obscure the policy identification of the equilibrium real exchange rate.

REFERENCES

- Balassa, Bela, "The Purchasing-Power-Parity Doctrine: A Reappraisal," *Journal of Political Economy*, December 1964, 72, 584-96.
- Bhagwati, Jagdish N., "Why are Services Cheaper in the Poor Countries?" *Economic Journal*, June 1984, 94, 279-86.
- Clague, Christopher, "Determinants of the National Price Level: Some Empirical Results," *Review of Economics and Statistics*, May 1986, 68, 320-3.
- Dornbusch, Rudiger, "Purchasing Power Parity," in John Eatwell, Murray Milgate, and Peter Newman, eds., *The New Palgrave Dictionary of Economics*, London: Macmillan, 1988, 1075-85.
- Hunter, Linda C. and Markusen, James R., "Per-Capita Income as a Determinant of Trade," in Robert C. Feenstra, ed., *Empirical Methods for International Trade*, Cambridge, MA: MIT Press, 1988, 89-109.
- Jones, Ronald W., "The Structure of Simple General Equilibrium Models," *Journal of Political Economy*, December 1965, 73, 557-72.
- Kravis, Irving B., Heston, Alan W. and Summers, Robert, *World Product and Income*, Baltimore: Johns Hopkins University Press, 1982.
- _____, _____, and _____, "The Share of Services in Economic Growth," in F. Gerard Adams and Bert G. Hickman, eds., *Global Econometrics*, Cambridge, MA: MIT Press, 1983, 188-218.
- _____, and Lipsey, Robert E., *Toward an Explanation of National Price Levels*, Princeton Studies in International Finance No. 52, Princeton, NJ: International Finance

- Center, Princeton University, 1983.
- ____ and _____, "The Assessment of National Price Levels," in Sven W. Arndt and J. David Richardson, eds., *Real-Financial Linkages Among Open Economies*, Cambridge, MA: MIT Press, 1987, 97-134.
- ____ and _____, "National Price Levels and the Prices of Tradeables and Non-tradeables," *American Economic Review*, May 1988 (*Papers and Proceedings*), 78, 474-8.
- Leamer, Edward E., *Sources of International Comparative Advantage*, Cambridge, MA: MIT Press, 1984.
- Linder, Staffan Burenstam, *An Essay on Trade and Transformation*, New York: Wiley, 1961.
- Neary, J. Peter, "Determinants of the Equilibrium Real Exchange Rate," *American Economic Review*, March 1988, 78, 210-5.
- Samuelson, Paul A., "Theoretical Notes on Trade Problems," *Review of Economics and Statistics*, May 1964, 46, 145-54.
- Thursby, Jerry G. and Thursby, Marie C., "Bilateral Trade Flows, the Linder Hypothesis, and Exchange Risk," *Review of Economics and Statistics*, August 1987, 69, 488-95.
- International Labour Organization, *Year Book of Labour Statistics*, Geneva, Switzerland: International Labour Office, 1979.

The Winner's Curse: Experiments with Buyers and with Sellers

By BARRY LIND AND CHARLES R. PLOTT*

This paper presents a replication and extension of experiments with the "winner's curse" which were initiated in John Kagel and Dan Levin (1984, 1986) and Douglas Dyer et al. (1989). The common-value auction involves firms bidding for an item of unknown common value. Since the value of the item is unknown, the winners can bid more than the value and thereby lose money. The winner's curse occurs if the winners of auctions systematically bid above the actual value of the objects and thereby systematically incur losses. The phenomenon is said to occur possibly in the bidding for such natural resources as mineral rights, where the value of the mineral is unknown but each firm has an estimate of the value. Due to the field nature of the data, doubts have existed as to the actual existence of the curse. The Kagel and Levin (1986) paper tested for the existence of the phenomenon in a laboratory setting. The hope is that, by achieving a thorough understanding of the phenomenon as it might exist in simple laboratory environments, economists will become better equipped to identify and study the phenomenon in more-complex field settings.

Kagel and Levin (1986) report the existence of a winner's curse, but as is the case with any seminal experiment, it is impossible to control for everything. After seeing their data and studying their experimental

procedures, one finds that there exist alternative explanations for what they saw. The winner's curse involves buyers who pay more than the value of an item and therefore experience a loss. Monetary losses in an experiment pose a problem because the experimenter generally has no means of collecting money from subjects. Subjects, knowing this, have reason to believe that the downside risk on their actions is truncated, and thus they might be prone to more risky actions than would be the case if they were forced to suffer full losses. In order to minimize this effect, subjects are frequently given a cash stake which they can lose. Kagel and Levin (1986) were aware of the problem, and they provided such a stake and used experienced subjects. They also required subjects to leave the experiment if and when the stake was lost. While these procedures provide some control, the possibility that losses could have contributed to the existence of the winner's curse has not been completely eliminated (Robert Hansen and John Lott, 1991). After a loss or two, a subject's reserve could be sufficiently low that prospective losses could exceed the balance. Thus, inflated bids carry no additional risk. Furthermore, one could theorize that experience with the curse facilitates learning and caution even in people who have had experience with bidding on other occasions. According to that theory, the process of removing bankrupt subjects succeeded in removing subjects less prone to the curse (i.e., those who had the experience of losing money and might adjust their behavior accordingly). Thus, subjects more prone to the curse would remain in the experiment and contribute to the existence of the curse. The situation is complicated even further by the possibility that subjects' beliefs about the reaction of other subjects to potential bankruptcy could cause general departures

*Barry Lind is a senior undergraduate and Charles Plott is a professor of economics at the California Institute of Technology. The financial support of the National Science Foundation and the Caltech Experimental Economics and Political Science Laboratory is gratefully acknowledged. We thank Michael Malcom for help and suggestions with the experimental design and Hsing-Yang Lee for his help with the graphics. We also thank Robert Hansen, John Kagel, Dan Levin, and John Lott for their comments.

from symmetric Nash-equilibrium behavior. Thus, a skeptic could claim that the existence, magnitude, and persistence of the winner's curse in the Kagel and Levin (1986) data were direct consequences of the way that Kagel and Levin's experimental procedures dealt with substantial losses by subjects. The technique used by Dyer et al. is the same as that used in Kagel and Levin (1986), so similar questions might be raised about it as well.

The strategy of the research reported here is to look for the phenomenon using procedures that avoid the bankruptcy problems. Two different sets of procedures are used. First, the "winner's curse" experiment in which subjects might lose money was conducted simultaneously with a second experiment in which subjects were making money. The second set of procedures involved competitors as sellers in a common-value auction. The winner's curse can appear in this setting as the sale of an item for less than it is actually worth to the seller. The seller's loss occurs as an opportunity cost only, so the possibility of bankruptcy does not exist.

The experiments using these different sets of experimental procedures produced several results which are the substance of the paper.

1. The winner's curse was observed in both experimental settings. In essence, the Kagel and Levin results were replicated.
2. The winner's curse observed by Kagel and Levin (1984, 1986) was not a consequence of their experimental procedures.
3. The winner's curse might diminish in size or frequency but does not completely dissipate over time.
4. The winner's curse is a general phenomenon exhibited by most agents.
5. Theories of "suboptimal" behavior advanced as explanations of the phenomenon do not explain the data as well as does the completely rational model in which the phenomenon does not exist at all theoretically.

The paper is organized as follows. In Section I, the experimental design is outlined. In Section II, some competing models are

discussed. Section III contains a statement of the measurement system. The results are in Section IV. The concluding section contains a discussion of conjectures that might advance an understanding of the phenomenon.

I. Experimental Design

The experiments were two types of common-value auctions. The first type of experiment was the common-value auction as conducted by Kagel and Levin (1984, 1986) in which buyers bid for an item of unknown value. Subjects agreed that if they suffered losses they would work them off at \$10 per hour.¹ In experiment 1, subjects participated in a sealed-bid private-value auction at the same time that they participated in a common-value auction in which the winner's curse might occur. In the second experiment, subjects participated in both a common-value auction in which they were buyers (experiment 2) and also in a common-value auction in which they were sellers (experiment 3). (In other words, experiments 2 and 3 were run simultaneously on the same subjects.) These secondary auctions constituted a source of funds which reduced the likelihood of bankruptcies in case the winner's curse was operative. These procedural changes were implemented so that subjects had full financial liability in the range of financial exposures that were likely to exist in the experiments.

The second type of experiments (experiments 3-5) were common-value auctions with competition among sellers as opposed to buyers. The sellers tendered offers to sell an item of unknown value. Each seller was given one item to sell. Their option was to keep the item and collect its value or sell the item and collect the revenues from the sale. The person with the lowest offer sold his item and received the asking price, while everyone else kept the item and received the value. In this common-value selling auc-

¹Only one subject suffered sufficient loss to be required to work. He worked about one hour to cover an \$8 loss.

tion, all subjects earned positive profits, including the winner, but the winner could suffer opportunity costs by selling the item for less than the amount received by those who did not sell the actual value of the item.

The experiments were conducted at the California Institute of Technology, using undergraduates as subjects. Most of the subjects had participated in other experiments prior to these and were familiar with the experimental environment. The subject pool serves as a partial control for the hypothesis that the curse might be due to confusion about instructions. The instructions read to the subjects are given in the Appendix. Prior to the experiment, the common values of the objects were determined by realization from a random-number table. Given the value of the object, "clues" or "signals" were drawn for each subject independently. Each subject was given a stack of slips of paper which contained the clues to the common value of the items being auctioned. The slips were stapled so that only the clue for the current period could be observed. The subject observed the clue and then submitted a bid. After the auction, all bids, signals, and the common value were posted. The winner was then announced. The subject removed the top slip to expose the clue for the next period.

The clue was called a signal about the true value of the item to be auctioned. The value of the item was randomly chosen from the range (x, \bar{x}) . If v was the item's value, then the signals were randomly chosen over an interval $(v - \varepsilon, v + \varepsilon)$, where ε is a positive value set by the experimenter. In order to avoid the winner's curse, the bidder must recognize that, if he wins and thus buys (sells) the object, then he probably has the highest (lowest) signal, which is probably above (below) the item's value. Therefore, in order for the person not to lose money, (forgo profits) he must bid (ask) significantly less (more) than this signal.

Five experiments were conducted. The first two were buyer markets which replicated one of the experimental settings of Kagel and Levin (1984, 1986). The next three were seller markets. All experiments were

conducted with seven subjects. Experiments 1 and 2 had the same set of predrawn signals, and experiments 4 and 5 had the same set of predrawn signals. The value of ε for the buying auctions was \$30, and it was 200 francs in the selling auctions. The range from which v was drawn was $(x, \bar{x}) = (\$25, \$225)$ for the buyer auctions and $(x, \bar{x}) = (150 \text{ francs}, 1,500 \text{ francs})$ for the seller auctions. (The franc values were \$0.0025, \$0.001, and \$0.0007 for experiments 3, 4, and 5, respectively.) The parameter choices reflect an attempt to identify unambiguously the curse, should it exist. The models reviewed below suggest that the curse becomes more severe with larger ε and a larger number of people. The parameters are those of Kagel and Levin (1986) that make the curse severe. Another consideration was cost. In the seller auctions, all subjects (except the seller) were paid the value of the item, which makes the experiments potentially expensive. For example, if the value in the seller auction had been drawn from the same distribution over dollars that it was drawn from in the buyer auction, then the cost of the experiment would have been on the order of \$875 ([expected value of v] $\times 7$) per period. The scaling factor that was chosen to reduce the cost keeps ε equal to the same proportion of the range of v and also permits many periods. This creates an obvious difference in marginal dollar stakes between the two types of experiments, with the potential "losses" due to departures from Nash behavior being very small in the selling experiment. Should otherwise inexplicable differences in behavior be observed, the magnitude of incentive would be an obvious line of research to pursue.

II. Models

Assume that v is drawn from a uniform distribution. Assume that each x_i is drawn independently from a uniform distribution over the interval $[v - \varepsilon, v + \varepsilon]$. If x_i is the signal observed by individual i and the structure is common knowledge, the theoretical problem is to model how i chooses a bid as a function of x_i .

At least four models make sense. The first is the *risk-neutral Nash-equilibrium model* of the associated bidding game.² The second model is based on the hypothesis that individuals make a specific type of calculation error but still conform to the general principles of game theory. We call this the *strategic-discounting model*. The third model is based on the hypothesis that people do not behave strategically. They only bid the expected value as if the situation were a simple second-price auction of a lottery and not one in which strategies might be important. This model is called the *naive model*. The fourth model, called the *private-value model*, postulates that individuals bid as if x_i were a private value of the object for each i . That is, individuals fail to understand the basic statistical relationship between value and signals.

The optimal bidding strategy according to the risk-neutral Nash-equilibrium model (RNNE) is to bid as a function of the signal (x_i). Under the buying auction, the optimal strategy is

$$(1) \quad b(x_i) = x_i - \varepsilon + Y$$

$$Y = [2\varepsilon / (n+1)]$$

$$\exp[-(n/2\varepsilon)(x_i - (\bar{x} + \varepsilon))]$$

where n is the number of subjects. Under the selling auction, the RNNE optimal strategy is

$$(2) \quad b(x_i) = x_i + \varepsilon - Y$$

$$Y = [2\varepsilon / (n+1)]$$

$$\exp[-(n/2\varepsilon)(-x_i + (\bar{x} - \varepsilon))].$$

A strategic-discounting model (SD) is postulated by Kagel and Levin (1986) for buyer auctions. The model is based on the hypothesis that individuals fail to recognize that the auction winner will be the subject with the highest signal. Kagel and Levin's strategic-discounting model can be general-

ized to the seller auction. The equations for the optimal bidding strategy under the assumption that the bidder fails to recognize that the winner has the highest (lowest) signal are

(3) buying auction:

$$b(x_i) = x_i - (2\varepsilon/n) + (Y/n)$$

(4) selling auction:

$$b(x_i) = x_i + (2\varepsilon/n) - (Y/n).$$

The above equations for the RNNE and SD models are only valid on the interval $\bar{x} + \varepsilon \leq x_i \leq \bar{x} - \varepsilon$.

The naive model (N) for both the buying auction and the selling auction simply has the bid equal to the signal. The bidding strategy for both types of auctions is

$$(5) \quad b(x_i) = x_i.$$

The final model, the private-value model (PV), holds that individual i makes the mistake of placing a private value x_i on the object and that the private value of each of the j others is independently drawn from the interval $x_i \pm \varepsilon$. By applying risk-neutral Nash theory to this situation, bidding functions can be derived. For buyers, the bidding function is

$$(6) \quad b(x_i) = x_i - \frac{\varepsilon}{n}$$

and for sellers it is

$$(7) \quad b(x_i) = x_i + \frac{\varepsilon}{n}.$$

III. Measurement Methodology

The four theoretical models lend themselves naturally to a single measurement system. The single regression for each individual,

$$(8) \quad b_{it} = \alpha_i + \beta x_{it} + \gamma Y_{it} + e_{it} \sim N(0, \sigma_i^2)$$

can be used as a measure of the accuracy of all four theoretical models. The equation

²Obviously, risk aversion is a natural extension. We have been unable to find a closed-form solution for the bidding functions.

TABLE 1—PARAMETER RESTRICTIONS IMPOSED BY COMPETING THEORETICAL MODELS

Model	Buying			Selling		
	α	β	γ	α	β	γ
Risk-neutral Nash equilibrium	-30	1	1	200	1	-1
Strategic discounting	-8.6	1	0.14	57.1	1	-0.14
Naive	0	1	0	0	1	0
Private value	-433	1	0	28.6	1	0

can be used for both buying auctions and selling auctions. A summary of the restrictions on the regression equation imposed by the competing theories is included as Table 1. As can be seen, all theoretical models predict $\beta = 1$. The intercept term can be interpreted as $\delta\epsilon$, where ϵ is the value of the range of the signal $\{x_i \in [v - \epsilon, v + \epsilon]\}$, so $\alpha = \pm 1\epsilon$ in the RNNE model, $\alpha = (2/n)\epsilon$ in the SD model, $\alpha = 0$ in the N model, and $\alpha = \pm(1/n)\epsilon$ in the PV model. For the RNNE model, γ is ± 1 ; it is $\pm 1/n$ for the SD model, and it is 0 for both the N and the PV models.

The measurement strategy is first to apply the unrestricted regression model. The coefficients can be compared to the theoretical values of the competing models. Then models with parameters as restricted by theory will be applied. The SSE of the unrestricted model can be used with the SSE of the restricted model to compute an F statistic (Chow test) for the hypothesis that the restrictions are not significantly different from the unrestricted measurements. The F statistic will also be used as a measure of the relative closeness of the competing models.

IV. Results

The results of primary interest bear on the existence of the winner's curse. Of secondary interest are results that might uncover the principles that govern individual decision behavior. The findings are summarized by five conclusions.

Conclusion 1: The winner's curse exists.

Evidence. The per-period profit from all auctions is used as a measure. In buying auctions, the profit is the actual value of the

object minus the purchase price of the auction winner. In selling auctions, the profit is the sale price of the object minus the actual price received by the winner. Thus, in selling auctions, a negative profit is an opportunity cost incurred because the item was sold for less than it was worth to the seller.

Table 2 lists the average per-period profits from all experiments. As can be seen, the winner suffers a loss in four of the five experiments on average. The table also reports the ratio of the number of auctions in which a loss occurred to the total number of auctions. In all cases, a large proportion of the auctions resulted in a loss. The only possible exception to the general tendency is experiment 4, which was characterized by a large number of attempts at collusion. In total, more than half of all auctions resulted in a loss.

Conclusion 2: The winner's curse persists with experience, but the magnitude and frequency of losses decline with experience.

Evidence. The frequencies of losses of the auction winners are divided into ten-period quartiles for every experiment in Table 3. The size of the average loss is also included in the table. As can be seen, the proportion of auctions in which losses occur is significantly greater than zero in all quartiles. Even after 20 or 30 auctions, the winners lose money more than 25 percent of the time. The frequency of losses decreases after the first 10 trials in all experiments except experiment 5.

The complete time-series of profits for experiment 5 is included as Figure 1. The figure also shows the profit that would have occurred if the agent had used the RNNE strategy. As can be seen, the winners' losses continue to occur even after 30 auctions.

TABLE 2—WINNER'S AVERAGE PROFIT AND THE LOSS FREQUENCIES
FOR ALL EXPERIMENTS
(PROFITS GIVEN IN DOLLARS, WITH FRANCS IN PARENTHESES)

Experiment	Average profit per period	Average RNNE predicted profit per period ^a	Number of periods with winner's loss (total number of periods)
1 (buyer)	-1.67 (-1.67)	11.13 (11.13)	12/20
2 (buyer)	-3.60 (-3.60)	8.85 (8.85)	10/17
3 (seller)	-0.022 (-8.88)	0.196 (78.44)	8/17
4 (seller)	0.021 (20.91)	0.069 (69.28)	13/35
5 (seller)	-0.013 (-18.55)	0.050 (70.85)	25/40

^aThe given RNNE equation is valid only for: $\bar{x} + \varepsilon \leq x_i \leq \bar{x} - \varepsilon$. Some of the winners' signals were not in this range, so no predicted RNNE profit is possible. Therefore, this average includes only periods for which the RNNE predicted profit can be calculated.

TABLE 3—FREQUENCY OF LOSSES FOR WINNERS IN ALL EXPERIMENTS
(PROFITS GIVEN IN DOLLARS, WITH FRANCS IN PARENTHESES)

Quartile	Experiment				
	1 (buyer)	2 (buyer)	3 (seller)	4 (seller)	5 (seller)
Periods 1-10					
Number of periods of loss	8/10	8/10	5/10	6/10	5/10
Average profit per period	-7.90 (-7.90)	-8.31 (-8.31)	-0.075 (-29.80)	-0.048 (-48.20)	0.001 (1.10)
Average RNNE profit per period ^a	4.53 (4.53)	5.70 (5.70)	0.177 (70.96)	0.060 (60.44)	0.048 (68.71)
Periods 11-20					
Number of periods of loss	4/10	2/7	3/7	2/10	7/10
Average profit per period	4.57 (4.57)	3.12 (3.12)	0.053 (21.00)	0.032 (31.60)	-0.016 (-22.40)
Average RNNE profit per period ^a	18.47 (18.47)	13.58 (13.58)	0.212 (84.85)	0.048 (48.15)	0.037 (52.68)
Periods 21-30					
Number of periods of loss				3/10	5/10
Average profit per period				0.058 (58.40)	-0.004 (-6.10)
Average RNNE profit per period ^a				0.104 (104.02)	0.090 (128.91)
Periods 31-40					
Number of periods of loss				2/5	8/10
Average profit per period				0.063 (62.80)	-0.033 (-46.80)
Average RNNE profit per period ^a				0.065 (65.34)	0.024 (33.72)

^aThe given RNNE equation is valid only for: $\bar{x} + \varepsilon \leq x_i \leq \bar{x} - \varepsilon$. Some of the winners' signals were not in this range, so no predicted RNNE profit is possible. Therefore, this average includes only periods for which the RNNE predicted profit can be calculated.

This experiment has a more severe curse than the other experiments. Unlike the other experiments, the frequency does not decline with experience.

The first two conclusions offer answers to the questions initially posed for experimental examination. The next series of conclusions reflect questions posed in an attempt

to understand why the phenomenon occurs. As was reviewed in the section above, only four theoretical models have been advanced. The first question posed was whether or not any of these four models represents the data in a statistical sense. Since the answer turns out to be negative, the next series of questions is an attempt to

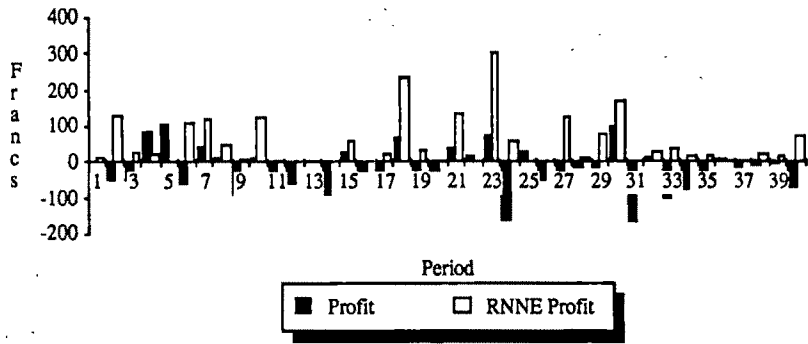


FIGURE 1. PER-PERIOD PROFIT AND RNNE PREDICTED PROFITS FOR EXPERIMENT 5

Note: In periods 5, 11, 13, 20, 22, 25, 26, 31, 36, and 37, the signal is not in the interval $[\bar{x} + \varepsilon \leq x_i \leq \bar{x} + \varepsilon]$, which is the valid range for the given RNNE function. Therefore, no RNNE predicted profits are shown

identify the “best” model and to ask why it fails.

Conclusion 3: All four models (RNNE, SD, N, and PV) can be rejected as statistical representations of the data.

Evidence. Table 4 contains the results of the Chow test described in the section above. In all cases, the statistical model with parameters as restricted by the competing theoretical models can be rejected as being significantly different from the unrestricted estimates. For example, the *F* statistic for rejecting the model at a 5-percent confidence is 2.64, while the statistic for the RNNE model for buying auctions is 30.53, and for selling auctions it is 5.93.

Conclusion 4: The RNNE model is the best model of the three considered, and the N model is the worst.

Evidence. The pooled data in Figure 2 show the relationship between individual signals and bids. The visual impression favors the RNNE model. The scattered data in the upper left of the figure for the seller auctions are the bids of a small number of subjects who were (evidently) signaling for collusion.

Table 5 contains the estimated coefficients from pooled data, which can be compared with the predictions in Table 1. With the exception of the x_i coefficient, β , the standard errors tell the same stories as do the Chow tests discussed below. The param-

TABLE 4—*F* STATISTICS FOR THE HYPOTHESIS THAT PREDICTIONS OF RESTRICTED REGRESSION AND UNRESTRICTED REGRESSIONS ARE THE SAME (DEGREES OF FREEDOM; 5-PERCENT *F* VALUES)

Model	Buying auctions	Selling auctions
RNNE	30.53 (3,226; 2.64)	5.93 (3,465; 2.61)
Strategic discounting	133.84 (3,226; 2.64)	91.22 (3,465; 2.61)
Naive	341.12 (3,226; 2.64)	160.13 (3,465; 2.61)
Private value	224.87 (3,226; 2.64)	122.99 (3,465; 2.61)

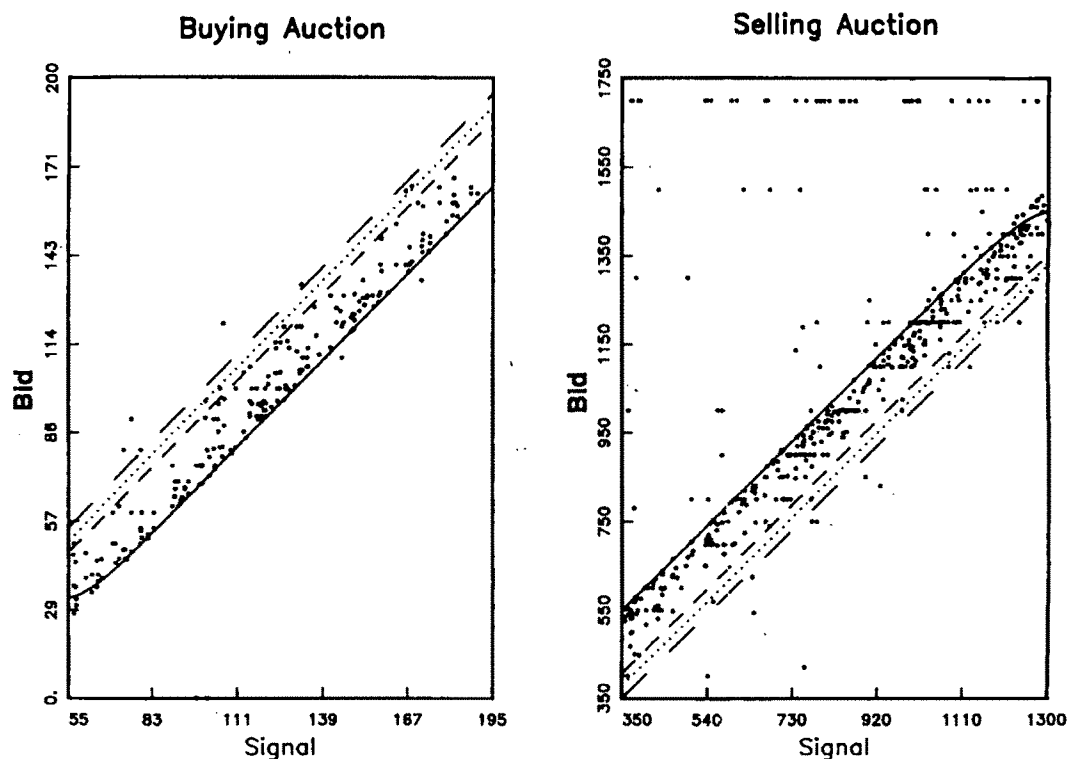


FIGURE 2. PAIRS OF SIGNALS AND BIDS FOR ALL INDIVIDUALS AND ALL EXPERIMENTS

Key: Longer-dashed line shows the prediction of the N model; shorter-dashed line shows the prediction of the SD model; dotted line shows the prediction of the PV model; solid line shows prediction of the RNNE model.

TABLE 5—ESTIMATED COEFFICIENTS FOR
POOLED INDIVIDUALS
(STANDARD ERRORS IN PARENTHESES)

Experiments	α	β	γ
1-2	-22.694 (3.156)	0.998 (0.024)	0.514 (0.674)
3-5	341.658 (38.335)	0.863 (0.046)	0.255 (1.585)

eter values estimated by the regression can be rejected as being equal to those of any of the four models. The β term is close to 1, but this is predicted by all models. The intercept term, α , is closest to that predicted by the RNNE model. The γ term has such a large standard error that little can be said other than that the sign is in

agreement with both the RNNE and the SD models.

The major support for the conclusion is simply a restatement of the F statistics in Table 4. If the F statistic is taken as a measure of accuracy, then the RNNE model is always more accurate than its closest competitor, the SD model. The PV model ranks third, and the N model is the worst. The F statistics for all models were also computed for each individual. For the 35 individual data sets, the RNNE model was the best fit (lowest F statistic) for 25, and 10 of these did not differ significantly from RNNE predictions. The SD model was best for all of the remaining 10 individuals, but in all cases, the data were significantly different from the predictions of the SD model.

TABLE 6—NUMBER OF TIMES WINNER HAD HIGHEST AND SECOND-HIGHEST SIGNAL

Experiment	Highest	Second-highest	Number of periods
1 (buyer)	14	3	20
2 (buyer)	9	8	17
3 (seller)	11	3	17
4 (seller)	21	8	35
5 (seller)	25	7	40

Conclusion 5: Failure of the RNNE model is not due to a few "irrational" people. Almost all agents experienced the "curse" and bid in a manner that was consistent with "curse" behavior.

Evidence. Table 6 gives the number of times that the winning bidder had the highest signal or the second-highest signal. The game-theoretic model predicts that the individual with the highest signal will win the auction. In each experiment, more than half of the auctions were won by the subject with the highest signal. As can be seen, decisions that resulted in winning the auction were not the result of some type of impulsive move by some agent with a lower signal; nor

was it the case that bids differed so much across subjects that the fundamental game-theoretic proposition that a positive relationship exists between bids and signals is destroyed. In fact, the empirical result in Table 5 that $\beta = 1$ is strong support for that part of the theory.

Table 7 gives the number of times each agent won the auction and the number of times each agent lost money as a result of winning the auction with a bid that was too high. As can be seen, the experience happens to most individuals. Of the 28 people who won two or more auctions, 20 of them lost money 50 percent of the time or more. Of the 35 subjects, only eight never lost money.

TABLE 7—NUMBER OF WINNING BIDS SUBMITTED AND NUMBER OF TIMES LOSSES OCCURRED, BY SUBJECT AND EXPERIMENT

Subject	Experiment				
	1 (20 periods)	2 (17 periods)	3 (17 periods)	4 (35 periods)	5 (40 periods)
1					
Number of winning bids	4	5	6	3	7
Number of times lost money	3	4	2	2	4
2					
Number of winning bids	3	1	1	6	7
Number of times lost money	2	0	0	3	4
3					
Number of winning bids	0	1	4	0	5
Number of times lost money	0	1	2	0	4
4					
Number of winning bids	5	2	1	5	4
Number of times lost money	3	2	0	2	2
5					
Number of winning bids	4	3	2	6	4
Number of times lost money	3	1	2	2	2
6					
Number of winning bids	2	2	2	12	7
Number of times lost money	0	0	2	3	4
7					
Number of winning bids	2	3	1	3	6
Number of times lost money	1	2	0	1	5

V. Closing Remarks

One question appears to be answered clearly: a winner's curse can be observed. A presumption exists about an answer to a second question: it appears that the curse can persist over many experiences. A major puzzle remains: of the models studied, the best is the risk-neutral Nash-equilibrium model, but that model predicts that the curse will not exist.

Part of the difficulty with further study stems from the lack of theory about the behavior of common-value auctions with risk aversion. Closed-form solutions which permit researchers to estimate models of "sub-rational" behavior have not been worked out. If the effect of risk aversion is to raise the bidding function as it does in private auctions, then risk aversion together with the strategic-discounting model might resolve the puzzle; but, of course, this is only a conjecture.

APPENDIX—INSTRUCTIONS

Instructions for buyer auctions are those that were used by Kagel and Levine (1986) and can be found in the appendix to their paper. Instructions were handed out to subjects, and all examples were also on the chalkboard. After the instructions were given to the subjects, they were read aloud by the experimenter, and then the following "test" was administered.

1. Buyer A gets a signal value of \$105.00. He bids \$100.00 but he is not the high bidder. His _____ (profit/loss) is \$ _____.
2. Buyer B gets a signal value of \$75.00. She bids \$60.00 and she is the high bidder. The value of the item is \$65.00. Her _____ (profit/loss) is \$ _____.
3. Buyer C gets a signal value of \$161.00. He bids \$132.00 and he is the high bidder. The value of the item is \$131.00. His _____ (profit/loss) is \$ _____.
4. Buyer D gets a signal value of \$120.00. The value of epsilon is \$30.00. Therefore, Buyer D knows that the value of the item is between \$ _____ and \$ _____.

Instructions for the Seller Auctions [Exact Transcript]

GENERAL

This is an experiment in the economics of market decision making. The instructions are simple and if you follow them carefully and make good decisions you might earn money which will be paid to you in cash.

In this experiment we will create a market in which you will act as sellers of a commodity in a sequence of trading periods. One unit of the commodity will be auctioned off in each trading period. There will be several trading periods.

Your task is to submit written asks for the commodity. The precise value of the commodity at the time you make your ask will be unknown to you. Instead, each of you will receive some information regarding the value of the commodity which you may find useful in determining your ask. The process of determining the value of the commodity and the information you receive will be described below.

The currency in these markets is francs. Each franc is worth \$ _____ to you.

The low ask gets the item and makes a profit equal to the ask. If you do not make the lowest ask on the item, you will earn the value of the commodity.

During each trading period, you will be selling in a market in which all of the other participants are also selling. After all asks have been handed in, all signals and asks will be posed on the blackboard. We will circle the low ask and post the value of the item.

The value of the auctioned commodity (V) will be assigned randomly and will lie between 150 and 1500 inclusively. For each auction, any value within this interval has an equally likely chance of being drawn. The value of the item can never be less than 150 nor more than 1500. The values V are determined randomly and independently from auction to auction. A high value of V in one period tells you nothing about the likely value in the next period. It does not even preclude the same value of V appearing in later periods.

Although you do not know the precise value of the item in any particular trading period, you will receive information which will narrow down the range of possible values. This will consist of a private information signal which is selected randomly from an interval whose lower bound is V minus epsilon, and whose upper bound is V plus epsilon. Any value within this interval has an equally likely chance of being drawn and being assigned to one of you as your private information signal.

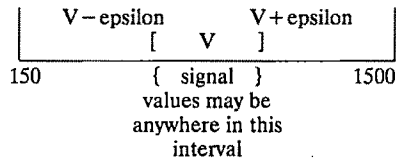
Throughout this experiment, the value of epsilon is 200.

PRIVATE INFORMATION SIGNALS

Although you do not know the precise value of the item in any particular trading period, you will receive information which will narrow down the range of possible values. This will consist of a private information signal which is selected randomly from an interval whose lower bound is V minus epsilon, and whose upper bound is V plus epsilon. ANY VALUE within this interval has an **EQUALLY LIKELY** chance of being drawn and being assigned to one of you as your private information signal. You will always know what the value of epsilon is.

For example, suppose that the value of the auctioned item is 762 and that epsilon is 200. Then each of you will receive a private information signal which will consist of a randomly drawn number that will be between 562 ($V - \text{epsilon} = 762 - 200$) and 962 ($V + \text{epsilon} = 762 + 200$). Any number in this interval has an equally likely chance to be drawn as your signal value.

The line diagram below shows what is going on in this example.



EXAMPLE

The value of the auctioned item is 762. This is the information each seller received, and the asks each seller made:

SELLER #	SIGNAL VALUE	ASK
1	590	703
2	756	900
3	838	947
4	634	778
5	716	775
6	847	920
7	642	825

In this example Seller #1 submitted the lowest bid, so he sells the item. His profit is the sale price 703. Seller #1 received 703 while the other sellers receive the value 762.

You will note that the value V of the auctioned item must always be between your signal value minus epsilon, and your signal value plus epsilon.

Finally, you may receive a signal value below 150 or above 1500. This merely indicates that the value V of the auctioned item is close to 150 or 1500.

Your signal values are strictly private information. **DO NOT REVEAL THEM TO ANYONE ELSE.** You are **NOT** to reveal your asks or profits, nor are you to speak to any other subject while the experiment is in progress.

You will not be told the value of V until after all the asks have been collected and posted.

No one may ask less than 0 for the item, nor may anyone ask more than 1700 (which is the maximum value of V plus epsilon). In case of ties for the low ask, we will flip a coin to decide who gets the item.

Are there any questions?

REFERENCES

- Dyer, Douglas, Kagel, John H. and Levin, Dan, "A Comparison of Naive and Experienced Bidders in Common Value Offer Auctions: A Laboratory Analysis," *The Economic Journal*, March 1989, 99, 108-15.
- Hansen, Robert G. and Lott, John R., Jr., "The Winner's Curse and Public Information in Common Value Auctions: Comment," *American Economic Review*, March 1991, 81, 347-62.
- Kagel, John H. and Levin, Dan, "First-Price, Sealed-Bid, Common Value Auctions: Some Initial Experimental Results," Center for Public Policy Discussion Paper 84-1, University of Houston, 1984.
- _____ and _____, "The Winner's Curse and Public Information in Common Value Auctions," *American Economic Review*, December 1986, 76, 894-920.

The Winner's Curse and Public Information in Common Value Auctions: Comment

By ROBERT G. HANSEN AND JOHN R. LOTT, JR.*

Do auction participants understand that auction winners tend to be those bidders who most overestimate the true value of an item? Papers by G. W. Gilley and G. V. Karels (1981), Kenneth Hendricks and Robert Porter (1988), and Stuart Thiel (1988) have presented real-world evidence that bidders act as if they understand the winner's curse. However, a recent series of papers in this *Review* and elsewhere by John Kagel and Dan Levin (1986), Douglas Dyer et al. (1989), and Kagel et al. (1989) report results from a series of sealed-bid laboratory experiments that provide evidence of individual "judgmental errors" and which lead the authors to question the general applicability of existing auction models.¹ This experimental evidence indicates that participants' behavior generally and significantly differed from what they specify as the

noncooperative (Nash) equilibrium bidding strategy for wealth-maximizing individuals in their laboratory context. More specifically, Kagel and Levin (1986) found that:

- 1) participants suffered a winner's curse in that bids often exceeded the conditional expected value of the item sold;
- 2) the high signal holder did not always win the auction;
- 3) reducing uncertainty over the item's value decreased the average price.²

One obvious question is: why does this experimental and real-world evidence produce such divergent results? In this paper, we point out that the experimental subjects in Kagel and Levin's research enjoyed limited liability for their losses and that this limited liability significantly affects the equilibrium bidding strategies for rational individuals. In fact, our goal is to show that limited liability in a common-value auction can explain each of these three judgmental errors in terms of perfectly rational behavior. Kagel and Levin misspecified the norm of rational equilibrium behavior, since they implicitly assumed unlimited liability, whereas the agents in their laboratory setting faced only limited liability.

The intuitive framework of our argument is easy to see. If a participant's cash balances during an auction series reached zero or went negative, he was no longer permitted to bid, and he was not held liable for

*Robert Hansen is an associate professor at the Amos Tuck School of Business at Dartmouth College and John Lott is a visiting assistant professor at the Anderson Graduate School of Management at the University of California at Los Angeles. We thank Gertrud Fremling and Bill Samuelson for reviewing an earlier draft of this paper, John Kagel for supplying us with the data, and seminar participants at Arizona State University, Cornell University, George Mason University, the University of Miami (Florida), Ohio State University, the Universities of California at Irvine and Los Angeles, Tulane University, the U.S. Securities and Exchange Commission, the U.S. Department of Justice, and the 1989 Public Choice/Economic Science Association Meetings. This paper was completed while Hansen was at the LEK Partnership and Lott was chief economist at the United States Sentencing Commission. Hansen received generous financial support from the Tuck Associates. Any remaining errors are our responsibility.

¹"Our experiments provide an empirical example of a market where individual judgment errors significantly alter market outcomes.... Although we reject the general applicability of Nash equilibrium bidding models..." (Kagel and Levin, 1986 p. 917). Their findings are already receiving wide attention (e.g., Colin Camerer, 1987; Richard Thaler, 1988).

²Finding 1 obviously contradicts wealth-maximizing behavior; finding 2 contradicts symmetric auction models, for in these models the high signal holder always wins; and finding 3 contradicts Paul Milgrom and Robert Weber's (1982) general theoretical result that reducing uncertainty will increase expected prices. (See James C. Cox and R. Mark Isaac (1984, 1986) for discussion on the definition of the winner's curse.)

that portion of any loss that drove his cash balances below zero; however, the auction rules stated (Kagel and Levin, 1986 p. 918) that bids in excess of one's cash balances were allowed.³ In bidding for an item of uncertain value, these cash balance rules give the subjects limited liability, in that bidders might incur losses larger than what they will have to pay. For example, in Kagel and Levin's (1986) auction series 3–8 it was possible for a bidder to receive a private information signal of \$80 even though the true value would be only \$50. If the bidder's cash balances still equalled his initial balances of \$10, a bid of anything over \$60 would make the individual's recorded loss greater than his actual loss.⁴

Lack of responsibility for large losses, at least for one-shot auctions, naturally leads to higher bidding. With unlimited liability, an increase in a bid of \$1 means that, conditional upon winning, profits are always \$1 less (losses are always \$1 greater) than without the bid increase; with limited liability, a bid increase of \$1 will mean no lower profits (no greater losses) for some states than before the bid increase. Thus, the cost of bidding higher is attenuated. This attenuation can be so strong that bids actually exceed the expected value of the item, when the expected value is calculated assuming unlimited liability. We show below that, as participants' cash balances approach zero, so that they have nothing to lose, bids approach the highest value that the item can take.

³It is of course important to note that a series of auctions was involved and that bankruptcy meant that an individual could not participate in subsequent auctions. As we point out in Section II, complete analysis of the effect that limited liability has on equilibrium bidding must account for multiperiod effects. Vernon Smith (1989 p. 160) uses the observation of bankruptcy to argue for another misspecification in Kagel and Levin's (1986) model. He argues that, while the number of bidders in their model is treated as fixed, it is actually endogenous.

⁴Kagel and Levin record profits/losses as: profits = (value of item) – (highest bid). Also, note that cash balances are \$10 for everyone only at the beginning of the auction series; profits and losses as the auctions proceed will cause balances to increase or decrease.

The experimental subjects, however, were not bidding in one-shot auctions. In a series of auctions, cash balances of the participants will not remain at their initial level; profits and losses of the participants cause cash balances to differ across participants and cause any one participant's balance to vary across auction periods. This observation leads to two further complications. First, asymmetry of cash balances leads to asymmetric bid strategies and therefore offers an explanation for result 2, that the high-signal holder will not always win the auction. As we establish below, those with low cash balances will bid higher than those with high cash balances, for the same information signal.

Second, the individual auctions within any one series cannot be considered independent. Bidding behavior in a prior auction affects the expected cash balances and optimal strategy for the next auction, so an optimal bid strategy for this prior auction must take future auctions into account. Taking this argument to the extreme, if a participant goes bankrupt in a period, he must leave the experiment and therefore forfeit the opportunity to bid in future periods. This multiperiod bankruptcy effect can, under certain circumstances, offset the one-period bid-increasing effect of limited liability. However, as we will show, the multiperiod nature of these auctions combined with limited liability in later periods can also cause bidders to take "unfair" gambles in prior periods, even though limited liability is of no use in that period itself. Given the intractability of this problem in a multiperiod context, one of the suggestions we have for future experimental research is that cash balances should be reinitialized prior to each auction in an auction series.

The following section initially ignores the multiperiod effects by focusing on one-period auctions with limited liability. This corresponds to the situation in the experimental studies, for which the theoretical analysis was limited to one-shot auctions with unlimited liability. Section II addresses the multiperiod aspects of limited liability and bankruptcy. Finally, in Section III, we test the importance of limited liability as an

explanation for this supposedly "irrational" behavior using the data employed by Kagel and Levin (1986). We also indicate certain fundamental design flaws in the experiments and what might be done to avoid these mistakes in the future.

I. A Simple Model

It would be useful to take the complete set of assumptions behind Kagel and Levin's (1986) experiments and derive the appropriate Nash equilibrium bid strategies. However, Kagel and Levin did not even solve for the complete equilibrium given unlimited liability, and their value and probability assumptions are not amenable to easy analysis.⁵ Limited liability, asymmetric cash balances, and a multiperiod context make their model even more intractable. For these reasons, we instead chose to analyze a simple discrete model in detail. The important results, however, are clearly (but only at considerable analytic expense) generalizable to more complex situations. Besides illustrating the effects of limited liability in Kagel and Levin's experiments, such a simple model of real-world auctions with limited liability should be interesting in and of itself. As we discuss below, one such application of the model is to auctions in which the seller guarantees the value.

Assume that two risk-neutral bidders compete in a one-shot auction for an item that has a value of either 0 or X . A statistical information service independently (conditional on item value) provides each bidder with one of two possible messages (message-1 and message-2) concerning item value. All calculations to follow will be based on the following probabilities:

prior probability: $\Pr(\text{value} = X) = p$;
posterior probability: $\Pr(\text{value} = X | \text{messages } i \text{ and } j) = p_{ij}$;

⁵The basic problem arises because of two conflicting assumptions: private information signals are uniformly distributed over $(X_0 - \varepsilon, X_0 + \varepsilon)$, where X_0 is the true value, but X_0 must be in (\underline{X}, \bar{X}) , the range of true values. Information signals less than $(\underline{X} + \varepsilon)$ or greater than $(\bar{X} - \varepsilon)$ are difficult to interpret because of endpoint problems.

conditional message probabilities: $\Pr(\text{opponent receives message-}i \text{ when message-}j \text{ received}) = m_{ij}$.

We further assume that message-1 is the "good" message, so that $p_{11} > p_{12} = p_{21} > p_{22}$.⁶

As a first step in the analysis, note that equilibrium bid strategies for the one-shot auction with unlimited liability are:⁷

$$(1) \quad b(m) = \begin{cases} \text{if } m = \text{message-2,} \\ b(m) = p_{22}X \\ \text{if } m = \text{message-1,} \\ \text{bid randomly in the interval} \\ (p_{22}X, m_{11}p_{11}X + m_{21}p_{22}X) \\ \text{according to the bid} \\ \text{distribution function} \\ F(b) = \frac{m_{21}(b - p_{22}X)}{m_{11}(p_{11}X - p_{22}X)}. \end{cases}$$

Derivation of this equilibrium follows from two observations: 1) contingent upon message-2, bidders must have zero expected profit and 2) contingent upon message-1, bidders must randomly make bids over some interval such that each bid yields the same expected profit. Note that bidding strategies are symmetric in this equilibrium, so that the high signal holder always wins the auction. Also, expected profits conditional on message-1 are positive; hence, the expected price must be less than the prior expected value of the item, pX .

The next step is to introduce starting cash balances and limited liability. Let c be each individual's initial cash balance and assume that the participant's liability is limited to c and that $c < p_{22}X$. Equilibrium with limited

⁶For simplicity, we have not derived these probabilities from the more fundamental probabilities of messages conditional on the state that occurs.

⁷Donald Hausch (1986) has derived the equilibrium for these assumptions for use in another context.

liability is then as follows:

(2) $b(m)$

$$= \begin{cases} \text{if } m = \text{message-2,} \\ b(m) = X - \left(\frac{1 - p_{22}}{p_{22}} \right) c \\ \text{if } m = \text{message-1,} \\ \text{bid randomly in the interval} \\ (X - [(1 - p_{22})/p_{22}]c, b^*) \\ \text{according to the bid} \\ \text{distribution function} \\ F(b) = \{m_{21}c[(p_{21} - p_{22})/p_{22}] \\ - m_{21}[p_{21}(X - b) - c(1 - p_{21})] \\ \times m_{11}[p_{11}(X - b) - c(1 - p_{11})]^{-1} \end{cases}$$

where

$$(3) \quad b^* = X - [c/(m_{11}p_{11} + m_{21}p_{21})] \\ \times [m_{21}[(p_{21} - p_{22})/p_{22}] \\ + m_{21}(1 - p_{21}) + m_{11}(1 - p_{11})].$$

As for the previous equilibrium, this bid function is derived by noting that expected profits conditional on message-2 must be zero and that profits conditional on message-1 must be the same for every bid included in the mixed strategy. The upper bound for the bid distribution is found by setting $F(b) = 1$. It is important to note that, as c decreases, the entire bid distribution moves toward the highest possible item value, X . Indeed, if $c = 0$, equilibrium calls for both bidders to bid X regardless of messages received. Thus, as bidders have less to lose, they become more aggressive in their bidding. Of course, as this occurs, bids will exceed the conditional expected value of the item.

Consider next what happens in the limited-liability case as uncertainty over item value is reduced. According to the general model of Milgrom and Weber (1982) and as is consistent with the unlimited-liability bidding equilibrium of Kagel and Levin, reducing uncertainty should increase expected

prices.⁸ An extreme example shows that this result need not hold under limited liability. Suppose that cash balances are arbitrarily close to zero for both participants, so that the equilibrium bids are also arbitrarily close to X no matter what the messages are. The expected price is therefore in the neighborhood of X . If we now remove all uncertainty, so that participants know whether an item's value is either 0 or X , bids will be either 0 or X , depending upon the message received.⁹ Expected price will then be the average of 0 and X , which is (considerably) less than the expected price with uncertainty and limited liability. More generally, with limited liability and less-extreme reductions in uncertainty, we expect the effect on expected prices to be ambiguous: the Milgrom and Weber effect works against the limited-liability effect. However, the limited-liability effect is stronger the lower are cash balances, so that the expected price will be more likely to fall with a reduction in uncertainty the lower are cash balances.

We have now established that Kagel and Levin's (1986) results 1 and 3 can be perfectly consistent with rational maximizing behavior. What remains to be shown is that, in a limited-liability context, the high signal holder might not always win the auction. The intuition is as follows: if cash balances are not equal across bidders, the equilibrium will also be asymmetric, with bids conditional upon the same message differing across participants. Bidders with low cash

⁸Notice that an extreme form of this result holds for our unlimited-liability model. With uncertainty, the expected price is less than prior expected item value; without uncertainty, expected price equals expected item value.

⁹With zero cash balances, equilibrium is arbitrary. With cash balances close to zero, the only equilibrium will be to bid 0 or X , depending upon the message received. Also, it is interesting to note that the zero-cash-balance equilibrium corresponds to a bidding equilibrium where the seller guarantees the value to be X . In that case, the seller always receives X as the high bid but must reimburse X if the value turns out to be zero. The expected price less the expected warranty is the prior expected value, pX . Giving a warranty is therefore similar, in terms of its revenue effects, to providing information.

balances will bid higher and win more auctions than participants with high cash balances.

To illustrate this formally, we take another extreme example: Suppose bidder A has a zero cash balance while bidder B has a cash balance of c . Then an equilibrium is as follows:

$$(4) \quad b_A = p_1 X + \alpha$$

$$(5) \quad b_B = p_1 X$$

where p_1 is the probability of the value equalling X conditional only on message-1 and α is the lowest bid increment. Given b_B , bidder A has no incentive to bid higher, since he always wins already; bidder A also has no incentive to bid any lower, for that would entail losing some profitable auctions. Given b_A , participant B would not bid higher (conditional on either message), because that would entail paying more than the expected value if he should win (which he would).

For this example, the zero-cash-balance participant always wins, even in the case in which he is not the high signal holder. In addition, the high-cash-balance bidder has zero expected profits; competing against someone who has nothing to lose cannot be profitable.

II. Multiperiod Effects

The main implication of a multiperiod setting is that equilibrium bidding strategies cannot be found for each separate auction by simply maximizing a bidder's expected profits for that auction alone. Instead, the objective for a participant must be to maximize the expected end-of-series cash balance, and the instruments available for achieving this objective will be the bids in all the auctions. To get an idea of the complexity that such analysis entails, note that an optimal first-auction strategy must take into account subsequent auctions. These subsequent auctions will necessarily involve asymmetric cash balances, and bidding behavior in the first auction will necessarily affect all participants' expected cash bal-

ances in subsequent auctions. Thus, to assess the profitability of bidding a little higher in the first auction, a bidder must account for how that will affect his and his competitors' cash balances and bid strategies conditioned on cash balances for later auctions. A simple example using two sequential first-price sealed-bid auctions can demonstrate the "irrational" implications of limited liability. Specifically, the example establishes that *limited liability can cause bidding in excess of expected value in initial auctions even if bankruptcy could never occur in that auction under unlimited-liability Nash equilibrium*.¹⁰

Suppose two risk-neutral bidders compete in a series of two auctions. In each auction, the item for sale has a value of either zero or \$20, with each event having a probability of 0.5. With unlimited liability, each bidder will bid the expected value of \$10 in each of the two auctions. (Ties are randomly broken.) Alternatively, suppose that each bidder begins the auction series with \$15 and that losses are limited to this amount. To see how bidding behavior changes with limited liability, suppose that bids remain at \$10 in the first auction. If the value turned out to be \$20 in the first auction, cash balances entering the second auction are \$15 and \$25, and each bidder will find it optimal to bid \$10 in the second auction. If, however, the value had turned out to be zero in the first auction, cash balances are \$15 and \$5 entering the second auction. Now the participant with a \$5 cash balance will be willing to bid more than \$10 in the second auction: a bid of \$10 gives him a 50-percent chance of ending with an expected cash balance of \$7.50 (either zero or \$15, with each equally likely) and a 50-percent chance of ending with his current bal-

¹⁰Prior reviews of our paper have focused on Kagel and Levin's initial auctions, in which cash balances are indeed high enough to make bankruptcy unlikely (but not impossible) under unlimited-liability equilibrium. The conclusion drawn is that limited liability is therefore unlikely to be a serious problem. However, as our example shows, the prospect of lower cash balances and bankruptcy in future auctions forces higher bidding even in the initial auctions.

ance of \$5. For example, a bid of \$10.01 gives him the certainty of ending with an expected cash balance of \$7.495. The low-cash-balance bidder effectively has a higher valuation and will win the second auction.

Was the initial assumption of \$10 bids in the first auction accurate? A simple analysis shows that bids of \$10 in the first auction imply expected end-of-series balances for each participant of \$15.625, more than they begin with. It is counterintuitive that profits should remain in a symmetric game such as this, and indeed, one can show that bids in excess of \$10 in the first auction are profitable. A bid of \$11.67 in the first auction eliminates any profit from the series as a whole (ending expected cash balances are \$15) and is in fact the equilibrium bid for each participant in the first auction.¹¹ Essentially, participants do not weigh losses in the first auction as heavily as they would under unlimited liability, because there is some chance that one dollar less carried into the second auction will not affect ending cash balances whatsoever (in bankruptcy, the seller takes all there is, be that \$5 or \$10).

To summarize, limited liability causes bidding in excess of expected values, even in auctions in which cash balances are initially high enough to eliminate the possibility of bankruptcy. (Interestingly, in our simple example, overbidding occurs only in the first auction, for in the second auction, the low-cash-balance participant wins by bidding $10 + \epsilon$.) It is therefore incorrect to treat the auctions of a series as being independent, and it is incorrect to argue that limited liability is not important overall on the basis of cash balances and bankruptcy probabilities in the initial auctions.

III. Reviewing the Evidence

Possibly Kagel and Levin's (1986) most important evidence is provided by their table 2, which implies overbidding in general and

a definite negative relationship between the number of bidders participating in an auction and the average "actual" profits from that auction. Their results, which are reproduced here in columns 2 and 3 in our Table 1, show their reported values for the average "actual" profits and the average "actual" profits as a percentage of profits that would occur under a risk-neutral Nash equilibrium (RNNE) by auction series. While Kagel and Levin use the term "actual" average profits, a more appropriate description is the *hypothetical* average profit that would have been obtained from the private-information bids had those actually been the effective bids. In auction series 3–8, Kagel and Levin employed a two-tier set of auctions in which a separate public-information auction was also performed after the private-information auction had been conducted. A coin flip determined whether the bids from the public- or the private-information auctions were the effective bids. If, as column 2 assumes, the bidders' private-information bids had always been the effective bids, several bidders who participated in all the auctions in experiments 3-large, 4, 5, and 8-large would have faced bankruptcy prior to the end of the auction series.

On this matter of limited liability, Kagel and Levin (1986 p. 896) state that, when an individual's "balances went negative, he was no longer permitted to bid; he was paid the \$4.00 participation fee and free to leave the experiment." In the actual auctions, individuals could not lose more than their initial starting balance of either \$8.00 (series 1 and 2) or \$10.00 (series 3–8). However, in calculating the hypothetical balances shown in columns 2 and 3, Kagel and Levin neither removed an individual when his hypothetical balances fell below zero nor truncated the bidder's losses at zero cash balances. They removed bidders from the sample only when the bidder's cash balances based on the effective bids reached zero, and not based on the hypothetical balances that they were calculating in column 2. Their accounting procedures resulted in larger recorded losses than it was possible for individuals ever actually to realize. Columns 4 and 5 then show what average profits would have been if losses had been stopped at

¹¹With a bid of \$11.67 in the first auction, expected cash balances will be either \$23.33, \$6.665, \$15, or \$15 at the end of the second auction, with equal chances of each.

TABLE 1—REVIEWING THE EVIDENCE ON AVERAGE PROFITS PER AUCTION WITH PRIVATE-INFORMATION BIDS

Auction series (number of periods) (1)	Kagel and Levin (1986)		Our calculations		Size of the last period's profit or loss (6)	Average profit up until last auction (standard deviation) ^a (7)	Percentage change in the last period's profits from (7) (8)	Significance level for test that (6) differs from (7) ^b (9)	Average profit up until last auction as percentage of RNNE (10)
	Average actual profits (2)	Percentage of RNNE (3)	Average actual profits (4)	Percentage of RNNE (5)					
5-7 Bidders:									
1 (18)	\$2.89	55.7	\$2.89	55.7	-\$3.75	+\$3.28 (3.64)	-214	0.04	62.6
3-large (11)	-\$2.92	-80.5	-\$2.09	-57.3	-\$10.00	-\$1.30 (5.42)	-669	0.07	-35.6
4 (25)	-\$0.23	-4.8	\$0.32	8.8	-\$9.85	+\$0.74 (5.95)	-1,431	0.05	20.3
5 (26)	-\$0.41	-7.8	-\$0.29	-5.3	-\$5.51	-\$0.076 (4.56)	-7,150	0.12	-0.1
7-large (18)	\$1.89	40.2	\$1.89	40.2	\$1.24	+\$1.93 (4.93)	-37	0.45	41.1
8-large (14)	-\$2.74	-54.8	-\$2.49	-49.5	-\$9.01	-\$1.98 (4.34)	-355	0.06	-39.4
Averages	-\$0.25		\$0.04		-\$6.15	+\$0.43	-1,643		
3-4 Bidders:									
2 (18)	\$4.61	92.6	\$4.60	92.2	\$10.98	+\$4.22 (4.32)	+160	0.07	84.6
3-small (14)	\$7.53	115.7	\$7.53	115.7	\$0.81	+\$8.04 (13.98)	-90	(0.31)	123.7
6 (31)	\$3.73	39.2	\$3.74	39.3	\$12.09	+\$3.46 (7.67)	+249	0.14	36.4
7-small (19)	\$5.83	68.1	\$5.83	68.0	\$8.32	+\$5.69 (7.79)	+46	0.37	66.4
8-small (23)	\$1.70	26.6	\$1.70	26.6	\$5.43	+\$1.53 (5.26)	+255	0.23	23.9
Averages	\$4.68		\$4.68		\$7.53	+\$4.58	+124		

^aBased on our calculation of actual balances.

^bThe null hypothesis is that the value from column 6 was drawn from the same distribution as those in column 7. Levels of significance for a one-tailed *t* test are shown.

zero cash balances based on the effective bids¹² and if losses or gains had been calculated in terms of both effective and non-effective private-information bids.¹³ While largely similar to Kagel and Levin's numbers, the average losses are reduced or eliminated for four of the auction series using 5-7 bidders (series 3-large, 4, 5, and 8-large), thus slightly weakening their results.¹⁴

¹²This seems to be the correct procedure because: 1) it is consistent with when bidders are removed from the auction and 2) it constrains their losses to what their true maximum losses can be. It is interesting to note that redoing Kagel and Levin's (1986) table 7 in a similar manner raises the hypothetical profits from public-information bids as a percentage of RNNE by about 6 percent.

¹³In doing these calculations, we limit losses to an individual's cash balances even when the bid is not an effective bid (because of the concomitant public-information auction). This seems to us to be more relevant than using bids that are not effective and not limiting losses, as Kagel and Levin (1986) do. An alternative is to limit losses and use only the effective private-information bids. Such a limit further increases profits for some auction series and further reduces the difference between the large- and small-numbers cases.

¹⁴We were able to replicate all of the results that Kagel and Levin (1986) reported in their table 2 and obtained very similar estimates to theirs in the other cases.

In examining Kagel and Levin's data, we noted another interesting aspect of their experiment. At first glance, the length of each auction series seems random, with experiments 7 and 8 lasting 37 periods and some, such as 1 and 2, lasting only 18 periods. However, as indicated in column 6 of Table 1, all five of the experiments involving a small number of bidders ended on a profit-making bid, while five of the six experiments involving a large number of bidders ended on a losing bid.¹⁵ Not only did the auction series that Kagel and Levin predicted should be profitable stop on winning bids and those experiments that they predicted would lose money stop on losses, but as column 8 indicates, they tended to stop when unusually large gains or losses occurred.

In the experiments involving a large number of bidders, the auction series ended with losses averaging 1,643 percent less than

¹⁵Even in the one case in which a series with a large number of bidders (7-large) was ended on a win, the percentage change in profits from the average wins up until the last auction was the smallest of any of the 11 changes shown in Table 1.

the average profits up to that point.¹⁶ The null hypothesis that each individual series with a large number of bidders (1, 3-large, 4, 5, 7-large, and 8-large) ended with an outcome drawn from a distribution with an average equal to the prior average was rejected at the 0.04, 0.07, 0.05, 0.12, 0.45, and 0.06 levels, respectively, in one-tailed t tests. We can strongly reject the hypothesis that all six series ended on an outcome equal to the prior average profits at the 0.0000004 level with a one-tailed t test.¹⁷ The experi-

¹⁶This bias can only be explained in one of the six cases with a large number of bidders (3-large) as resulting from one of the bidders going bankrupt so that there was no longer the requisite number of bidders to qualify as a case with a large number of bidders. In that instance, when the fifth bidder went bankrupt, the large-number-of-bidders auctions series ended, and the auction series with a small number of bidders started. The experiments should have been set up so that the series were never constrained to end on a loss. However, this does not explain why the other auctions with large or small numbers of bidders ended in the systematic manner that they did (see footnote 18 for a related discussion).

The bias obviously cannot be explained by other variables that were under the experimenters' control, such as by any systematic variations in ϵ . In experiment 7-large, ϵ was lower at the end of the auction series than at the beginning; in auction series 1, 2, and 3-small, ϵ started and ended on the same values, and in auction series 5 and 6, ϵ was declining at the end of the auction series. In the remaining auctions, ϵ was increasing at the end of the auction series.

Equally important, the second-to-last period's profits or losses do not suggest any systematic variations. In four of the auctions with small numbers of bidders, the second-to-last period's profits were less than the average profits up until that last auction, and in two of the cases, the reverse was true. In the auctions with large numbers of bidders, the second-to-last period's profits were less than the average profits up to that point in three cases, and in the other two cases, the reverse was true. This provides some evidence against claims that there were some underlying trends in the series.

¹⁷An interesting feature of the data for the series with a large number of bidders was that auction series 1 ended on the largest loss of the 18 auctions in that series, auction series 3-large ended on the largest loss of its 11 auctions, auction series 4 ended on the second-largest loss of its 25 auctions, auction series 5 ended on the second largest loss of its 26 auctions, auction series 8-large ended on the second largest loss of its 14 auctions, and auction series 7-large (the only one to end on a profit-making bid) ended on the ninth-smallest value out of 19 auctions.

ments involving a small number of bidders ended with a profit that was averaging 124 percent higher than the average profit in all the preceding auctions.¹⁸

The importance of this can be seen by looking at column 7 in Table 1. Dropping the results from the last auction in each series alters the previous findings, with the average profits rising for all of the auctions with large numbers of bidders and falling for all but one of the auctions with small numbers of bidders. *The average profits per auction for the auctions with large numbers of bidders changes from -25 cents in Kagel and Levin's results to +43 cents here.*¹⁹ While the average profits as a percentage of RNNE are still lower for the cases with large numbers of bidders, the relationship is no longer as systematic as claimed by Kagel and Levin and largely owes itself to three of the 11 series: 3-small, 3-large, and 8-large. While acknowledging that removing those three series from the sample may be discarding valuable information, it is interesting to note that the average profits as a percentage of RNNE for the remaining four series involving 3-4 bidders is 52.8, and it is 31 for the four remaining series for 5-7 bidders.²⁰

¹⁸The series with a small number of bidders showed a similar, though much less pronounced, pattern to that noted in footnote 15; auction series 2 ended on the third-largest gain of the 18 auctions in that series, auction series 6 ended on the sixth-largest gain of its 31 auctions, auction series 7-small ended on the eighth-largest gain of its 19 auctions, and auction series 8-small ended on the fifth-largest gain of its 23 auctions. However, auction series 3-small, which already had the highest average profits, ended on the eight-largest gain of its 14 auctions.

¹⁹Since most of the truncations of the losses occurred during the final auction period, most of the change between columns 2 and 7 would have occurred even if we had not done the transformation shown in column 4.

²⁰Kagel has informed us that the rule used to determine when to end each auction series was that each series would last two hours. A possible explanation for the above pattern might be that this foreseen last period caused the bidders to behave in a systematically biased manner, depending upon whether they were part of a large or a small group of bidders. Why these systematically different effects exist is unclear. Since the same participants are often used in one long experi-

While the preceding indicates that Kagel and Levin's evidence is not quite as strong as they claim, the most important question still remains: to what extent can the overall results be explained by limited liability? In order to answer this question, we reran the specification shown in their table 4 across the entire sample.²¹ (For the moment, we will not try to control for the bias over when the auction series were ended.) Kagel and Levin (1986 p. 899) argue that the bid should be a function of the private signal, the dispersion in possible values that the private signal can take from the true value (ϵ), the number of bidders, and an adjustment for how the bid should vary in accordance with the risk-neutral Nash equilibrium (Y).

Fortunately for our purposes, it was also possible to obtain the level of each bidder's cash balances prior to the bids. Our earlier discussion suggests that, if limited liability is important, bids should be negatively related to cash balances. Those who face small losses because of constrained cash balances should increase their bids. The potential importance of cash balances may be quite

large, since cash balances are less than ϵ for 80 percent of the 1,101 bids and the median is \$6.84 (or 38 percent) less than the median value of ϵ . Importantly, theory also suggests that cash balances should affect bidding differentially according to the private signal received (note that in the equilibrium of Section II, the effect of cash balances varies with the message). Thus, we have included an interactive term between a bidder's cash balances and his private-information signal.

In addition, there was one other source of problems in Kagel and Levin's experimental setup. While the bidders knew the minimum and maximum possible values that the true value could equal, the private signal was less than the minimum value that the true value could take in 35 cases and was greater than the maximum value in 25 cases. Obviously, those who know that their private signal is greater than the maximum possible true value should discount their bid (as compared to the RNNE), while those who know the reverse is true will increase it. In addition, in those cases, ϵ could provide a biased estimate of the risk associated with the bid.²² We sought to control for these effects by using two dummy variables to indicate when the private signal is outside these possible values.

Using ordinary least squares, the reestimated regressions from Kagel and Levin's table 4 after controlling for the level of cash balances prior to each bid are

$$\begin{aligned}
 (6) \quad & \text{private-information bid} \\
 & = 0.9809(\text{private signal}) \\
 & \quad (438.92) \\
 & + 0.2592(\text{number of bidders}) \\
 & \quad (1.5151) \\
 & - 0.5644 \epsilon \\
 & \quad (17.13) \\
 & - 0.07311(\text{cash balances prior to bid}) \\
 & \quad (2.512) \\
 & - 0.16 \\
 & \quad (0.13)
 \end{aligned}$$

iment involving both small and large number of bidder auctions, even if the bidders know when the entire auction will end, they can only tell when one of the two types of auctions will end. To put it differently, auction series 7 lasts 37 periods, of which the first 19 are for the case with a small number of bidders. All six bidders employed in groups of four during the first 19 auctions continue to be employed altogether in a group of six for the remaining 18 auctions. Even if these bidders genuinely face some type of last-period effect at the 37th auction, why would the bidders perceive the 19th auction as distinct from the previous 18 auctions? Whatever the reason, any evidence of the winner's curse depends upon the inclusion of these last periods in the auction series.

²¹Unlike Kagel and Levin (1986), we have run our regressions over the entire data set. Kagel and Levin's justification for excluding the data near the extreme values that the item can take is that their bidding equilibrium only holds in the interior range. Since we believe that their bidding model is incorrect, we also believe that it would be incorrect to delete observations on the basis of that model. Our regressions are consistent with their table 2, which provides their strongest evidence, in which they do not exclude these observations. We also ran all of our regressions only over the same observations employed by Kagel and Levin, but this did not qualitatively alter our findings with respect to the cash-balance variables.

²²Because of this truncation, large values of ϵ will not seem to elicit very large reductions in the price due to risk, simply because the true value of ϵ is not as large as that measured.

($R^2 = 0.994$, sum of squared residuals = 62,129.3, $N = 1,101$) and

(7) private-information bid

$$\begin{aligned}
 &= 0.984 (\text{private signal}) \\
 &\quad (446.48) \\
 &+ 0.1367 (\text{number of bidders}) \\
 &\quad (0.851) \\
 &- 0.6074 \varepsilon \\
 &\quad (19.55) \\
 &+ 0.014 Y \\
 &\quad (12.38) \\
 &- 0.0773 (\text{cash balances prior to bid}) \\
 &\quad (2.83) \\
 &- 0.73 \\
 &\quad (0.63)
 \end{aligned}$$

($R^2 = 0.995$, sum of squared residuals = 54,505.9, $N = 1,101$), where the absolute t statistics are shown in parentheses. Two features of these results immediately stand out. First, the individual bidders' cash balances are always significantly negatively related to the level of the bid, and second, inclusion of the cash balances significantly reduces both the size of the coefficient for the number of bidders and its significance.

When an interactive term for cash balances and the private signal is added in these two specifications, the net effect of cash balances on the size of the bid is negative for values of the private signal greater than 75.7 and 60. These critical values are only about one-half the private-information signal's mean value of 136.3. While the coefficient on the number of bidders is always positive and significant at the 0.10 level in single-tailed t tests when we do not control for cash balances, it is significant at that level in only two of the four regressions that employ cash balances.²³ In-

clusion of the cash-balance terms reduces the coefficients on the number-of-bidders variable by between 10 percent and 44 percent, with an average decline of 25 percent.

Not only is the variable for cash balances statistically significant, it is also economically important, at least compared with the number-of-bidders variable. A one-standard-deviation increase in a bidder's cash balance (\$8.26) implies that the average bid will decline by between 42 and 64 cents using equations (6) and (7). Similarly, a one-standard-deviation increase in the number of bidders (1.37) implies an increase in the size of the bid of between 19 and 36 cents. To put it in slightly different terms, using equation (7), the number of bidders would have to increase from 3 to 7.68 in order for the number-of-bidders variable to offset a one-standard-deviation increase in cash balances (note that the number of bidders in the data only ranges from 3 to 7).

When the dummy variables for whether the private signals were greater than the maximum possible true value or less than the minimum possible value are included, their effects are quite significant. The coefficients for those dummy variables indicate that the bids are typically decreased by almost \$4 when the private signal is greater than the maximum possible true value and increased by almost \$23 when the reverse is true. The inclusion of these two terms increases the significance and the size of the coefficients for ε and the number of bidders and greatly reduces the significance and size of Y .

As noted earlier, the periods on which the auctions series were ended differ significantly from prior periods. The preceding regressions were then reestimated without those last periods and are reported in Table 2. The results coincide closely to those reported above in all but two respects. The most obvious differences is that the coefficients for the number of bidders are both from 10-percent to 30-percent smaller and less likely to differ significantly from 0, once those last periods are removed from the sample. In only four of the eight regressions is the coefficient significant at the 0.10 level in a single-tailed t test, and that is true for

²³Exclusion of the intercept term in the regressions shown in Table 2 results in the number-of-bidders variable being significant at the 0.10 level for a single-tailed t test in only two of the five regressions that control for cash balances. Excluding the intercept, however, has almost no effect on the other variables.

TABLE 2—REGRESSIONS EXCLUDING THE LAST AUCTION IN EACH SERIES: ENDOGENOUS VARIABLE IS THE LEVEL OF PRIVATE-INFORMATION BIDS
(ABSOLUTE *t* STATISTICS IN PARENTHESES) (*N* = 1,046)

Specification	Private signal	Number of bidders	Cash-balance prior to bid	Cash-balance \times private-signal interaction	ϵ	<i>Y</i>	Private signal > maximum value possible	Private signal < minimum value possible	Constant	<i>R</i> ² [sum of squared residuals]
1	0.9822 (437.1)	0.2865 (1.682)	—	—	-0.5750 (17.49)	—	—	—	-1.35 (1.2)	0.994 [59,390.7]
2	0.9817 (434.5)	0.2081 (1.193)	-0.0627 (2.009)	—	-0.5571 (16.37)	—	—	—	-0.29 (0.2)	0.995 [59,161.2]
3	0.9851 (462.8)	0.1840 (1.147)	—	—	-0.6214 (19.93)	0.0014 (11.68)	—	—	-0.52 (0.5)	0.995 [52,513.8]
4	0.9846 (460.6)	0.0975 (0.593)	-0.0689 (2.349)	—	-0.6019 (18.68)	0.0014 (11.74)	—	—	-0.64 (0.5)	0.995 [52,236.6]
5	0.9920 (520.2)	0.2766 (1.973)	—	—	-0.6720 (24.53)	0.0004 (3.36)	-4.005 (3.16)	22.891 (17.79)	-1.50 (1.5)	0.996 [39,952.4]
6	0.9917 (516.2)	0.2264 (1.573)	-0.0400 (1.541)	—	-0.6605 (23.28)	0.0004 (3.43)	-3.960 (3.12)	22.770 (17.68)	-0.78 (0.8)	0.996 [39,861.2]
7	1.0050 (185.8)	0.2837 (1.635)	0.1426 (2.694)	-0.0018 (4.77)	-0.5430 (16.07)	—	—	—	-3.70 (1.3)	0.995 [57,892.6]
8	1.0030 (196.3)	0.1599 (0.976)	0.0913 (1.821)	-0.0014 (3.92)	-0.5896 (18.34)	0.1360 (11.38)	—	—	-2.1 (1.5)	0.995 [51,473.5]

only two of the five cases when we control for cash balances. In fact, the removal of the intercept term for the regressions shown in specifications 7 and 8 in Table 2 actually causes the coefficients on the number of bidders to switch signs and equal -0.04 and -0.19. The results in Table 2 also diverge from the earlier ones in that the coefficients for cash balances (when they appear alone—specifications 2, 4, and 6), while they are always still significant and negative, are associated with larger *P* values (greater probability of type-I error) than they were previously. The net effect of cash balances is still negative at relatively small values of the private-information signals of \$79 and \$65 for specifications 7 and 8 in Table 2.

We also noted in Table 1 that the results there were sensitive to the inclusion of a few of the auction series. If our discussion of cash balances is correct, the exclusion of series 3-large and 8-large should weaken the importance of the cash-balance term, since bidders in those two auctions suffered unusually large losses and thus faced greater reductions in liability in the preceding auctions. Rerunning the specifications from Table 2, excluding series 3-large and 8-large (but without dropping the last auctions in the remaining series), the coefficients for the cash-balance terms were still always significant and negative, but were less so than for the cases shown in Table 2.

Even more interesting, however, was the effect that this change had on the coefficient for the number-of-bidders variable. The coefficient was now negative in seven of the eight specifications and was negative and significant at the 0.15 level for a single-tailed *t* test in five cases. The *t* statistic for the single case in which the coefficient was positive was only 0.068. These results are important, since a negative coefficient is exactly what the risk-neutral Nash equilibrium predicts. Since the likelihood that a bidder will suffer the winner's curse increases with the number of bidders that he outbids, bidders facing a large number of bidders adjust their bids downward accordingly. The fact that their results do not hold over 9 of the 11 auction series at least indicates that some caution is necessary in accepting any claims of the winner's curse being a function of the number of bidders.

As an alternative to controlling for cash balances, we replaced cash balances with a variable equal to ϵ minus cash balances, but with all negative values set equal to zero. The logic here is that if cash balances were greater than ϵ and individuals bid the expected value of an item, they would not create a limited-liability problem. Therefore, running cash balance by itself should bias downward the effect of the coefficient, since we do not expect cash balances greater than ϵ to have any effect on the level of the

bids. These changes left largely unaltered both the sizes and significance levels of the cash-balance effects previously discussed.²⁴

One general objection to our empirical work is that, despite using the bidder's cash balances immediately preceding his bid, the causation does not run from low cash balances to higher bids, but instead, individuals who bid too high tend to have low cash balances. In order to control for these individual-specific characteristics, which may account for the earlier importance of the cash-balance effect, we reran the earlier specifications shown in Table 2 (though over the entire sample) including a dummy variable to denote each of the individual bidders (bidder number 7 in auction series 8-large is represented by the intercept). The corresponding regressions (shown in Table 3) produce somewhat mixed results. While the effect of cash balances on the level of bids is always negative, the coefficients in specifications 2 and 4 in Table 3 are significant at only the 0.127 and 0.109 levels, respectively, for a one-tailed *t* test. In specifications 5 and 6 of Table 3, where the interaction between cash balances and private signals is allowed, the cash-balance and interaction coefficients are both highly significant. The net effect of cash balances becomes negative at values of the private signal greater than 88 and 92. Also, the coefficients for the number-of-bidders variable range from only 37 percent to 93 percent of the size of those found by Kagel and Levin and are significant at the 0.10 level for a one-tailed *t* test in four of the six regressions.²⁵

Finally, Kagel and Levin's table 2 suggests that the bidder with the highest private signal only wins the auction 71 percent of the time. The question is whether know-

ing what the lowest bidder's cash balances were will help explain the probability that he will win the auction. To test this, we ran a dummy variable indicating which bidder won the auction in each series on two new variables: 1) the difference between the bidder with highest private signal in each auction and the private signal of each bidder in that series ("difference in an auction's private signals") and 2) the difference between the bidder with the highest cash balance in each auction and the cash balance of each bidder in that series ("difference in an auction's cash balances"). (These variables are zero for the bidders with the highest private signal or cash balances in any particular auction.) However, as demonstrated in Tables 2 and 3, it is not just the difference in cash balances that is relevant, but also their level. Therefore, we estimated separate regressions by multiplying the "difference in an auction's cash balances" variable in each successive regression with a different dummy variable according to whether the cash balances were less than \$2, \$4, \$6, or \$8.

The results (shown in the regressions in Table 4) provide some support for our hypothesis. The Probit regressions show that the effect of the "difference in an auction's private signal" variable is consistently significant. However, only the interaction term for "difference in an auction's cash balances \times the dummy for when cash balances are less than \$2 is significant at the 0.10 level for a single-tailed *t* test.²⁶ One possible explanation for the cash-balance variable's relatively low levels of significance is that when cash balances are low for any individual bidder in an auction they tend to be low

²⁴ We also tried using a variable equal to cash balances divided by ϵ and obtained similar results.

²⁵ We note here another effect possibly confounded with the effect of cash balances. Suppose an individual loses money in an initial auction and therefore bids high in the following auctions (in accordance with his low cash balance). The cash-balance effect therefore becomes confounded with the individual dummy variable. This multicollinearity should reduce the significance level of the cash-balance variable.

²⁶ While the effects of the interaction term are not statistically significant, for cash values less than \$4, \$6, and \$8, the *t* statistics are successively smaller. Regressing the dummy variable for the auction's winner on the "difference in an auction's private signals" and "difference in an auction's cash balances" without the interaction effect reveals that the effect of the private signal is relatively unchanged but that the effect of cash balances is only significant at the 0.125 level for a one-tailed *t* test. However, this indicates that continuing the interaction terms for still higher levels of cash balances would eventually increase the level of significance.

TABLE 3—REGRESSIONS TAKING ACCOUNT OF THE INDIVIDUAL CHARACTERISTICS OF BIDDERS:
ENDOGENOUS VARIABLE IS THE LEVEL OF PRIVATE-INFORMATION BIDS.
(ABSOLUTE *t* STATISTICS IN PARENTHESES) (*N* = 1,101)

Specification	Private signal	Number of bidders	Cash balance prior to bid	Cash-balance \times private-signal interaction	ϵ	<i>Y</i>	Constant	Adjusted <i>R</i> ² [sum of squared residuals]
1	0.9669 (410.0)	0.5097 (2.2095)	—	—	-0.6416 (20.01)	—	5.3248 (2.7766)	0.996 [44,123.7]
2	0.9670 (409.5)	0.5145 (2.2304)	-0.0363 (1.1483)	—	-0.6303 (18.62)	—	5.4317 (2.8285)	0.996 [44,077.7]
3	0.9716 (436.3)	0.2577 (1.1945)	—	—	-0.6719 (22.43)	0.0013 (12.72)	6.0127 (3.3653)	0.996 [38,226.5]
4	0.9717 (435.9)	0.2629 (1.2186)	-0.0389 (1.2392)	—	-0.6541 (20.81)	0.0013 (12.73)	6.1311 (3.4276)	0.996 [38,170.6]
5	0.9873 (196.1)	0.6070 (2.6465)	0.1317 (2.8581)	-0.00149 (4.5376)	-0.6243 (18.60)	—	1.9536 (0.9524)	0.996 [37,669.9]
6	0.9872 (210.0)	0.3410 (1.5828)	0.1052 (2.1170)	-0.00115 (3.7285)	-0.6541 (20.81)	0.0013 (12.42)	3.4330 (1.7885)	0.996 [39,861.2]

for all the other bidders as well and thus do not greatly affect the relative ranks of bids within any given auction. After the highest cash balance in each auction is removed, the average value for the "difference in auction's cash balance" variable is only \$8.49.²⁷

²⁷There is also evidence that the values of ϵ were not independent of the auction results. At first glance, the changes in ϵ seem to be random with $\epsilon = 12$ lasting anywhere from 6 to 18 auction periods, depending on the auction series; likewise $\epsilon = 30$ ranges from 6 to 12 periods, and $\epsilon = 18$ ranges from 8 to 16 periods. The problem is that if ϵ tends to be relatively large for the cases with large numbers of bidders, limited liability would tend to be more of a problem for these cases, and it would appear that those bidders were behaving relatively less "rationally." Likewise, if ϵ only tended to be low when cash balances were also low and tended to be high when cash balances were high, the limited-liability problem would be present a larger percentage of the time. To examine this, we regressed ϵ on the 2^K set of variables for the bidder's cash balances prior to that auction, the square of that value, and a dummy variable which equalled 1 when the number of bidders was greater than or equal to 5. The coefficients consistently showed that both more bidders and higher values of cash balances were significantly related to increased values of ϵ . For example, we obtained

$$\begin{aligned} \epsilon = & 1.77(\text{cash balances}) \\ & (38.4) \\ & + 2.71(\text{number of bidders dummy}) \\ & (5.80) \\ & - 0.0318(\text{cash balances squared}) \\ & (23.60) \end{aligned}$$

(sum of the squared residuals = 63,715.3, *N* = 1,101),

where the absolute *t* statistics are shown in parentheses. In the cases with small numbers of bidders, increased cash balances are associated with increased values of ϵ up until \$27.65, which is 2.5 times greater than the median value of cash balances of \$11.16; ϵ is also \$2.71 higher when there are at least five bidders. Since the experimenters determined both the number of bidders and ϵ , questions of causality do not seem to be important for that relationship. The problem of causality between ϵ and cash balances is more serious. One partial answer is that the level of cash balances is determined from the previous auction's bid and thus is known before ϵ is changed. However, Granger causality tests with a single lag—an admittedly imperfect solution—did not reject either hypothesis for the direction of causation. Given the possibility that ϵ was endogenous to cash balances and the number of bidders, we reran the regression specifications shown in Table 2 (though employing the entire sample) and the specification shown above in this footnote using two-stage least squares. While the variables for cash balances always had effects of the predicted sign and were significant at lower levels (smaller probability of type-I error) than those previously reported, the coefficients for the numbers of bidders variable were only significant in those regressions corresponding to specifications 1 and 5 in Table 2. For example, rerunning equations (6) and (7) resulted in

private-information bid

$$\begin{aligned} = & 0.977(\text{private signal}) \\ & (406.3) \\ & - 0.0057(\text{number of bidders}) \\ & (0.029) \\ & - 0.1167\epsilon \\ & (1.698) \\ & - 0.1490(\text{cash balances prior to bid}) \\ & (3.6484) \\ & - 5.5 \\ & (4.33) \end{aligned}$$

TABLE 4—REGRESSIONS EXPLAINING THE PROBABILITY OF WINNING AN AUCTION:
ENDOGENOUS VARIABLE IS A DUMMY VARIABLE FOR THE AUCTION'S WINNER
(ABSOLUTE *t* STATISTICS IN PARENTHESES) (*N* = 1,101)

Difference in an auction's private signals	Difference in an auction's cash balances multiplied by a dummy = 1 when balances < \$2	Similar to (2) but multiplied by a dummy = 1 when balances < \$4	Similar to (2) but multiplied by a dummy = 1 when balances < \$6	Similar to (2) but multiplied by a dummy = 1 when balances < \$8	Constant	Log likelihood [average likelihood]
-0.06364 (11.66)	0.06695 (1.2989)	—	—	—	-0.2664 (4.295)	-448.69 [0.6653]
-0.06371 (11.67)	—	0.04508 (1.0955)	—	—	-0.2648 (4.268)	-449.08 [0.6651]
-0.006389 (11.71)	—	—	0.00592 (0.4024)	—	-0.2628 (4.201)	-449.52 [0.6648]
-0.06386 (11.69)	—	—	—	0.003896 (0.3518)	-0.2642 (4.153)	-449.53 [0.6648]

Overall, our evidence indicates that any proof of the existence of the winner's curse and its positive variance with the number of bidders is much weaker than suggested by Kagel and Levin, whose results are sensitive to 1) accounting procedures that resulted in larger recorded losses than it was possible for individuals ever actually to realize; 2) the ending of auction series for large numbers of bidders on significantly lower profits

than for other prior auctions; 3) not controlling for the bidder's level of cash balances; and 4) the inclusion of only two of the 11 auction series. The evidence also offers additional support for our limited-liability explanation by showing that the level of cash balances affected who would win an auction.

IV. Conclusion and a Suggestion for Future Research

Two alternative explanations have been advanced for the experimental subjects' behavior in Kagel and Levin's (1986) study. Kagel and Levin argue that the behavior can only be a result of judgment errors—a failure by bidders to appreciate the fact that auctions tend to select as winners those who most overestimated item value. We have argued that the observed behavior may be a perfectly rational response to limited liability and low cash balances. The experimental data do not reject the hypothesis that bidders are responding rationally to limited liability. As to whether or not bidder's behavior still significantly differs from the norm of a rational equilibrium, resolution of that question using the current data awaits derivation of the correct norm. A possibly more feasible alternative to calculating this correct standard is to rerun the experiments in terms of a series of one-shot auctions,

($R^2 = 0.993$, sum of squared residuals = 62,650.7, $N = 1,101$) and

private-information bid
= 0.979 (private signal)
(423.0)

-0.108 (number of bidders)
(0.569)

-0.127 ϵ + 0.0012 Y
(1.927) (9.315)

-0.0773 (cash balances prior to bid)
(3.994)

-5.5
(4.33)

($R^2 = 0.994$, sum of squared residuals = 66,492.3, $N = 1,101$), where the absolute *t* statistics are shown in parentheses. The removal of the last auction in each series leaves the cash-balance variable unchanged but results in the number-of-bidders variables being negative and significant at the 0.10 level in a single-tailed *t* test.

after each of which the cash balances are reinitialized. Such an experimental setting would allow for an accurate test of the one-shot common-value auction models that are analyzed in the theoretical literature.²⁸

²⁸In their response to our criticism, Kagel and Levin cite Barry Lind and Charles R. Plott (1991). We agree that Lind and Plott's methodology of using "seller auctions" is an innovative way to control for limited-liability problems. We would like to point out, however, that Lind and Plott's results are actually consistent with our hypothesis that limited liability has exacerbated findings of the winner's curse in experimental auctions. Specifically, Lind and Plott's results show a greater winner's curse in the buyer auctions, and the RNNE predictions perform relatively better for the seller auctions. For example, in using table 2 from Lind and Plott (1991), we calculate actual average profits of -\$2.56 for the buyer auctions (the RNNE prediction is \$10.08) and actual average profits of -\$0.002 for the seller auctions (the RNNE prediction is \$0.08). (See Glenn Harrison [1989] for a discussion of the problem of not giving the participants in experiments sufficient incentive to make the correct calculations.) Thus, actual profits are 125 percent below those predicted for the buyer auctions but only 102 percent below those predicted for the seller auctions. Also, 59 percent of the buyer auctions had losses, whereas only 50 percent of the seller auctions had losses. Turning to the statistical tests in Lind and Plott, the critical *F* statistic for rejecting the RNNE restrictions at the 5-percent level is 2.64; the actual *F* statistic for the buyer auction is 30.53, while for the seller auctions it is only 5.93. Put differently, the observed significance level is much lower for the seller auctions than for the buyer auctions. From a Bayesian viewpoint, the seller auctions, which are not contained by limited liability, present weaker evidence against the RNNE prediction. This is exactly our point.

REFERENCES

- Camerer, Colin F., "Do Biases in Probability Judgment Matter in Markets? Experimental Evidence," *American Economic Review*, December 1987, 77, 981-97.
- Cox, James C. and Isaac, R. Mark, "In Search of the Winner's Curse," *Economic Inquiry*, October 1984, 22, 579-92.
- _____, and _____, "In Search of the Winner's Curse: Reply," *Economic Inquiry*, July 1986, 24, 517-20.
- Dyer, Douglas, Kagel, John H. and Levin, Dan, "A Comparison of Naive and Experienced Bidders in Common Value Offer Auctions: A Laboratory Analysis," *Economic Journal*, March 1989, 99, 108-15.
- Hausch, Donald B., "Multi-Object Auctions: Sequential vs. Simultaneous Sales," *Management Science*, December 1986, 32, 1599-611.
- Hendricks, Kenneth and Porter, Robert H., "An Empirical Study of an Auction with Asymmetric Information," *American Economic Review*, December 1988, 78, 865-83.
- Gilley, G. W. and Karels, G. V., "The Competitive Effect in Bonus Bidding: New Evidence," *Bell Journal of Economics*, Autumn 1981, 12, 637-49.
- Harrison, Glenn, "Theory and Misbehavior of First-Price Auctions," *American Economic Review*, September 1989, 79, 749-62.
- Kagel, John H. and Levin, Dan, "The Winner's Curse and Public Information in Common Value Auctions," *American Economic Review*, December 1986, 76, 894-920.
- _____, _____, Battalio, Raymond and Meyer, Donald J., "First-Price Common Value Auctions: Bidder Behavior and the Winner's Curse," *Economic Inquiry*, April 1989, 27, 241-58.
- Lind, Barry and Plott, Charles R., "The Winner's Curse: Experiments with Buyers and with Sellers," *American Economic Review*, March 1991, 81, 335-46.
- Milgrom, Paul R. and Weber, Robert J., "A Theory of Auction and Competitive Bidding," *Econometrica*, September 1982, 50, 1089-122.
- Smith, Vernon, "Theory, Experiment and Economics," *Journal of Economic Perspectives*, Winter 1989, 3, 151-69.
- Thaler, Richard H., "Anomalies," *Journal of Economic Perspectives*, Winter 1982, 2, 191-202.
- Thiel, Stuart E., "Some Evidence on the Winner's Curse," *American Economic Review*, December 1988, 78, 884-95.

The Winner's Curse and Public Information in Common Value Auctions: Reply

By JOHN H. KAGEL AND DAN LEVIN*

In their comment, Robert Hansen and John Lott (1991 p. 360) (hereafter HL) argue that the violations of Nash-equilibrium bidding theory reported in Kagel and Levin (1986) (hereafter KL) and related publications (Kagel et al., 1989) "*may be a perfectly rational response to limited liability and low cash balances*" [emphasis added].¹ While HL's examples emphasize the importance of not overlooking the potential effects of limited liability on bidding, we controlled for this problem in two ways: 1) by choosing a design that minimizes the incentive to overbid relative to the Nash equilibrium as a result of limited-liability considerations and 2) by providing subjects with cash balances that were large enough that, for almost all bidders (48 out of 50), balances were always high enough that it *never* paid to deviate unilaterally from the Nash-equilibrium bidding strategy of a single-shot game.

The protection our experimental design affords against overbidding as a result of

limited liability is easily seen through clarification of HL's own first example. HL consider a bidder with a private-information signal of \$80 in an auction where the value of the item is \$50, the bidder has a cash balance of \$10, and $\varepsilon = \$30$ (the interval around the true value from which private-information signals were drawn). They note that any bid over \$60 would make the individual's loss greater than his cash balance, concluding that the lack of responsibility for large losses is likely to lead to higher bidding. What HL fail to point out (or do not realize) is that, in this example, the risk-neutral Nash-equilibrium (RNNE) bid is \$52.27 in a market with four bidders, or \$50.41 in a market with seven bidders, so that the bidder is fully liable for all losses (and a good deal more), relative to the Nash-equilibrium bid.² As long as bidders have sufficient cash balances to cover their maximum losses relative to their Nash-equilibrium bids, their overbidding *cannot* be rationalized in terms of a Nash equilibrium.³ Rather, overbidding must be explained on some other grounds, such as the judgmental error underlying the winner's curse.

Although HL's analysis of our data shows a statistically significant cash-balance effect, we show that the regression models HL employ are misspecified. Once their specification errors are corrected, we find 1) a statistically significant positive time trend in bidding, which is the exact opposite effect of what HL's multiperiod example predicts but

*Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260, and Department of Economics, University of Houston, Houston, TX 77204. Research support from the Information Science and Technology Division and the Economics Division of the National Science Foundation, the Alfred P. Sloan Foundation, the Russell Sage Foundation and the Energy Laboratory of the University of Houston is gratefully acknowledged. We thank John Ham for helpful discussions of the statistical analysis and Susan Garvin for research assistance. We retain full responsibility for any errors.

¹HL's opening statement would have the reader believe that the winner's curse is a phenomenon identified only in the laboratory but not present in "real world" markets. In fact, the term winner's curse was first introduced in the literature by petroleum geologists (E. C. Capen et al., 1971) in order to characterize what they believed to have been the outcome of bidding on outer-continental-shelf (OCS) oil-lease sales for the period 1954-1969 (also see the references cited in KL [p. 894] and Wilson's [1990] review of field data on the winner's curse).

²Here, we assume that $\bar{x} = \$25$, as HL do (see Section I for the equilibrium bid function).

³Literally, this is only a local result. However, in the simulations reported in Section I, we find that in auctions with four or seven bidders, with $\varepsilon = \$30$ and cash balances of \$4.50 (which 48 out of the 50 bidders always had), even larger deviations from the RNNE bid function are not profitable.

is consistent with our argument that multiperiod effects may promote less aggressive bidding, thereby mitigating the winner's curse, and 2) a statistically significant cash-balance effect indicating that each \$1 increase in balances reduced bids \$0.08. While the sign of the cash-balance coefficient is negative, our analytic results and simulations show that this cannot be the result of limited-liability forces. This interpretation is supported through a regression specification that eliminates all subjects who went bankrupt (including the two whose cash balances were low enough that they might have been motivated to deviate from the Nash equilibrium), with virtually no effect on the sign, the size, or the statistical significance of the cash-balance coefficient. Finally, results from another common-value auction experiment (Barry Lind and Charles Plott, 1991), which replicates our design while eliminating the limited-liability problem, reproduces our primary results, providing additional evidence that limited-liability forces do not account for the winner's-curse phenomenon reported in our experiment.

I. Limited Liability and Nash-Equilibrium Bidding in the KL Design

HL's argument that limited liability for losses may result in overbidding and other (apparent) violations of Nash-equilibrium bidding theory for perfectly rational bidders is well taken.⁴ Indeed, the central procedural issue to be resolved in designing our experiment was what to do about losses, since if bidders were not held accountable for them, it might promote bidding above the RNNE. On the other hand, if bidders were to be held liable for their losses, there was a need to specify "bankruptcy" conditions and to devise procedures to maintain a constant number of bidders. As a result, we designed an experiment that makes HL's examples (the effects on bidding when cash

balances are zero and the multiperiod effects of bidding in auctions with zero expected profit) irrelevant. Following HL, we consider single-shot auctions first and then move on to issues of limited liability in repeated auctions.

A. Single-Shot Auctions

Our experiment involved only experienced bidders who had experienced the (more) extreme effects of overbidding found among inexperienced bidders in common-value auctions (Kagel et al., 1989). Under our design, the RNNE produced positive expected profits that were increasing in x_0 , the value of the auctioned item, and in the x_i 's (individual bidders' private-information signals), which were uniformly distributed on the interval $[x_0 - \varepsilon, x_0 + \varepsilon]$. Bidders started the experiment with a cash balance of \$8 or \$10 which, they knew from experience, was theirs to keep and add to through bidding below x_0 or to lose as a result of bidding above x_0 . Further, looking at the subjects' cash balances, of the 50 bidders in our eight auction series, only two were ever bidding with cash balances below \$4.50 (for a little over half the bidders, cash balances never fell below starting levels).⁵ That is, all bidders started out being liable for losses up to \$8–\$10, and virtually all of our bidders were always liable for the first \$4.50 in losses suffered. This has important implications for the applicability of HL's limited-liability argument to our experiment.

In our design, for signals in the interval $\underline{x} + \varepsilon < x_0 < \bar{x} + \varepsilon$, the RNNE bid function is $b(x) = x - \varepsilon + Y$, where $Y = [2\varepsilon / (N + 1)] \exp[-N(2/\varepsilon)(x - (\underline{x} + \varepsilon))]$ and N stands for the number of active bidders in the market.⁶ Since Y contains a negative

⁵Most bankruptcies occurred for subjects whose cash balances stood at \$4.50 or above.

⁶HL (p. 349) argue that KL "...did not even solve for the complete equilibrium given unlimited liability." This is not correct (see footnote 8 in KL). What we do not have is an analytic solution for the differential equation specifying the equilibrium bid function for $x_i > \bar{x} - \varepsilon$, which constitutes only a small portion of the signals' domain. Further, contrary to HL's assertions

⁴We define a winner's curse as having occurred when (i) there is a high positive correlation between bids and signals and (ii) there is bidding above the expected value conditional on winning the auction.

exponential it diminishes rapidly as x_i moves beyond $\bar{x} + \varepsilon$. Further, since a bid of $\bar{x} - \varepsilon$ protects against *all* losses (x_i are drawn from the interval $[x_0 - \varepsilon, x_0 + \varepsilon]$), as long as bidders have sufficient cash balances to cover Y , they are fully liable for all losses relative to the Nash-equilibrium bid. The maximum possible value for Y is $2\varepsilon/(N+1)$ so that with $\varepsilon = \$12, \18 , and $\$30$ (the values employed in the experiment) maximum losses were $\$4.80, \7.20 , and $\$12.00$, respectively, with $N = 4$ and $\$3.00, \4.50 , and $\$7.50$, respectively, with $N = 7$.⁷ (The seeming paradox of limited liability having a greater potential impact with $N = 4$ than with $N = 7$ is explained by the fact that the RNNE bid function is decreasing in N .) Consequently, given the starting cash balances subjects had at their disposal, limited liability could only have been a factor in explaining early deviations from the RNNE in experiments with $N = 4$ and $\varepsilon = \$30$. Further, with cash balances of $\$4.50$ or more, which virtually all bidders always had, the limited-liability argument can only explain overly aggressive bidding with $N = 4$ and with $\varepsilon = \$30$ for $N = 7$. However, in our experiment (see table 3 in KL), the winner's curse was most prominent with $N = 7$ and $\varepsilon = \$12$ and $\$18$, as these treatment conditions produced negative average profits (all others produced positive average profits). This is exactly the opposite of what one would predict on the basis of HL's limited-liability argument!

Simulations help clarify the impact of limited liability in our design. In particular, they indicate that, with a cash balance of $\$4.50$, the limited-liability argument is not relevant to experiments with $N = 4$ or $N = 7$ for values of ε up to $\$30$. Table 1 reports

the results of these simulations. We consider expected profits of bidder 1 for unilateral deviations from the Nash strategy, assuming that all other bidders follow the Nash strategy. The simulation employed 25,000 draws of x_0 and corresponding random draws of x_i about x_0 , with liability for losses limited to $\$4.50$. Expected profits were calculated for player 1 at the RNNE bid and for unilateral increases in player 1's bid above the RNNE (with all other bidders playing the Nash strategy), employing increments of $\$0.50$ through the bid of $\$3.00$ and $\$1.00$ increments thereafter.

The first part of Table 1 shows the effect on player 1's expected profit with $\varepsilon = \$30$, given that $\bar{x} < x_0 < \bar{x} + \varepsilon$ and x_0 is distributed uniformly on the interval $[\bar{x}, \bar{x} + \varepsilon] = [\$30, \$230]$ (with $\bar{x} < x_0 < \bar{x} + \varepsilon$, we are sampling from the interval with the lowest expected profit, as there is zero expected profit for $\bar{x} < x_i < \bar{x} + \varepsilon$, with expected profit becoming positive and increasing monotonically for $x_i > \bar{x} + \varepsilon$). Note that unilateral deviations from the RNNE *reduce* total profit, as bidding above the RNNE lowers profit in those auctions that would have been won anyway, and these losses are larger than the increased profits earned from winning more auctions. To be sure, as HL argue, limited liability attenuates the cost of bidding above the RNNE here, as the reduction in profits for any given deviation is smaller than in the case of unlimited liability. Nevertheless, profits are reduced. Note also that expected profit is higher with $N = 7$ than with $N = 4$, as the RNNE calls for less aggressive bidding with larger N for signal values in this interval.

The second part of Table 1 shows the corresponding simulations for $\varepsilon = \$30$ and $\$115 < x_0 < \145 , the midpoint of possible x_0 values, assuming that x_0 is distributed uniformly on the interval $[\bar{x}, \bar{x} + \varepsilon] = [\$30, \$230]$. There is positive expected profit under the RNNE for all signal values in this interval. Here too, deviations from the RNNE produce uniformly lower expected profit. Further, as long as the bid deviation is $\$4.50$ or less, there is no difference between expected profit with and without unlimited liability, since in our design a bid of $x_i - \varepsilon$,

there is no difficulty in interpretation or error in our design when $x_i < \bar{x}$ or $x_i > \bar{x}$; it is just that, given the known bounds on x_0 and the symmetric distribution of signal values around x_0 , these signals are more informative about x_0 than x_i farther from the end points.

⁷Maximum potential losses occur with $x_i = \bar{x} + \varepsilon$ and $x_0 = \bar{x}$. For x_i at the midpoint of the interval of possible values used in our typical auction series ($x_i = \$125$, $\bar{x} = \$25$, $\varepsilon = \$225$), maximum losses would be $\$0.11$ for $N = 4$ and less than $\$0.005$ for $N = 7$.

TABLE 1—THE EFFECT OF UNILATERAL DEVIATIONS FROM RNNE WITH LIABILITY FOR LOSSES LIMITED TO \$4.50 AND $\epsilon = \$30$

Deviation from RNNE (dollars)	$\underline{x} \leq x_0 \leq \bar{x} + \epsilon$				$115 \leq x_0 \leq 145$			
	$N = 4$		$N = 7$		$N = 4$		$N = 7$	
	Number of auctions won	Average profit	Number of auctions won	Average profit	Number of auctions won	Average profit	Number of auctions won	Average profit
0.00	6,321	0.2487	3,638	0.3418	6,319	3.007	3,636	1.093
0.50	7,020	0.1979	4,423	0.3274	6,511	2.988	3,837	1.086
1.00	7,764	0.1467	5,263	0.3124	6,697	2.965	4,040	1.078
1.50	8,567	0.1000	6,128	0.2857	6,915	2.956	4,261	1.076
2.00	9,373	0.0388	7,058	0.2537	7,114	2.931	4,471	1.059
2.50	10,158	-0.0492	8,005	0.2084	7,315	2.913	4,677	1.039
3.00	10,989	-0.1173	9,016	0.1674	7,505	2.875	4,871	1.009
4.00	12,654	-0.2820	11,083	0.0484	7,974	2.855	5,301	0.960
5.00	14,314	-0.4745	13,167	-0.1453	8,388	2.772	5,726	0.886
6.00	15,887	-0.7150	15,054	-0.3839	8,844	2.743	6,145	0.817

Note: Profit was averaged over all 25,000 auctions in the sample.

which closely approximates the RNNE bid function, insures against losses.

Finally, our simulations confirm the fact that, if a player has a cash balance of \$0.00, unilateral deviations from the RNNE increase expected profit, as HL's argument suggests (this happens when $\underline{x} < x_0 < \bar{x} + \epsilon$ and for values of x_0 immediately beyond $\bar{x} + \epsilon$, where Y is still significant). However, as we will show below, this does not necessarily imply that bidders with low cash balances will bid more aggressively than the RNNE model predicts.

B. Multiperiod Auctions

HL also discuss the effect of limited liability in multiperiod auctions, as it affects single-period outcomes. They note correctly that, under certain circumstances, the threat of bankruptcy from overly aggressive bidding in multiperiod auctions can completely offset the limited-liability forces that may promote bidding above the RNNE in a single-period auction. HL do not specify these circumstances but, instead, provide an example of a multiperiod auction in which limited liability in later periods induces bidders to take "unfair" gambles in early periods, even though cash balances are sufficiently large to cover all potential losses in these early periods. Unlike their single-

period-auction examples, in this example there is zero expected profit in each auction period, as there are no private-information signals, only common knowledge concerning the underlying distribution of x_0 .

In contrast, our design offers significant positive expected profits at the Nash equilibrium. Thus, bankruptcy can be quite costly in terms of precluding future positive profits. This force works *against* any single-period limited-liability forces promoting more aggressive bidding. In addition, given our experimental design, it suggests that more aggressive bidding ought to be observed as time goes on. HL have conveniently switched examples, from auctions with positive expected profit to auctions with zero expected profit, in going from their single-period to multiperiod examples, rendering their multiperiod example irrelevant to our design.

II. Reviewing the Evidence in KL

In recalculating the per-period profit from our experiment, HL note a tendency for auctions with small numbers of bidders to end with gains and for those with large numbers of bidders to end with losses; they also note that all auctions stopped with above-average gains or losses. HL conduct t tests demonstrating that these last-period

gains and losses were larger than average. They wonder how these effects could arise when the length of each auction series seems random.

The explanation is quite simple. Our experimental design called for increasing ε over time, starting with smaller ε 's, which would limit losses, to give subjects time for learning and refreshing their memories, ending with larger ε 's, for which expected profits would be larger (see table 1 in KL). In auctions with small numbers of bidders, where bids were consistently below the expected value conditional on winning, these increases in ε resulted in above-average profits for later auction periods. For auctions with large numbers of bidders, in which bids were often above the expected value conditional on winning, increases in ε would result in larger losses in these later periods. The latter did not happen on average, however, as bidders shaved their bids as ε increased (see table 3 in KL).⁸ Nevertheless, the variance of expected profit conditional on winning was increasing in ε . Further, there is an endogeneity problem in auctions with large numbers of bidders that HL fail to account for. When bidders went bankrupt, N decreased, at least to the extent that we did not have substitutes to replace the bankrupt bidders. This happened in auction series 3-large and 8-large, two of the worst offenders in HL's calculations.⁹ Bankruptcies always involved losses and usually involved relatively large losses.

HL report a series of regressions identifying cash-balance effects on bids in our private-information auctions. Most of their statistical analysis employs ordinary-least-squares techniques, which is inappropriate since, as HL themselves note, it does not account for individual-subject differences, so

that the direction of causality may run from individuals who bid too high and thus tend to have low cash balances, rather than from low cash balances to higher bids. This problem is corrected in HL's table 3, where they employ fixed-effects regression models, producing "somewhat mixed results" (p. 358) compared to their OLS regressions. These regressions are still flawed, however, as HL completely ignore the endpoint effects of signal values near \underline{x} and \bar{x} , and they fail to explore (or to report their explorations of) any multiperiod effects present in the data.¹⁰ Our regression analysis, reported in Table 2, corrects these deficiencies.

First, our regressions, unlike HL's, are restricted to signal values in the interval $\underline{x} + \varepsilon < x_i < \bar{x} - \varepsilon$, as was done in our original paper, since the form of the RNNE bid function is quite different for x_i outside this interval. Along with individual-subject dummy variables, all of our specifications include a subject $\times \varepsilon$ interaction term permitting each subject to respond differently to changes in ε , which proves to be statistically significant under all specifications (HL do not account for this interaction effect). Our preferred specification is reported as regression 1 in Table 2. It includes a time-trend variable (normalized auction period) to capture the effects of any multiperiod auction forces (HL have no time-trend variable in their regressions) and a cash-balance variable (cash balances at the time the bid was made).¹¹ The cash-balance variable is negative and statistically significant, suggest-

⁸In KL, we conjecture that this results from risk aversion.

⁹Auctions series 3-small is defined as the periods after the last bankruptcy in auction series 3. In auction series 8, one subject went bankrupt approximately 15 minutes before our two-hour time period was up, reducing the number of bidders from seven to six. Rather than proceed for one or two periods with $N=6$, the auction series was terminated.

¹⁰HL (p. 355) suggest that signal values near these endpoints were "...one other source of problems in Kagel and Levin's experimental setup." Nothing could be further from the truth (see footnote 6). Failure to account for the additional information associated with these signal values does pose statistical problems however, which HL did not account for in their table 3.

¹¹The auction-period variable is normalized to account for the different numbers of auction periods in the different auction series. Subjects were recruited for a fixed period of time, and we typically conducted as many auction periods as time permitted. Similar results are obtained using an untransformed period variable. Subject dummies are specified for each subject in each auction series; so, a subject participating in more than one auction series will have a separate dummy variable in each series.

TABLE 2—REGRESSION ANALYSES SHOWING THE IMPORTANCE OF CASH BALANCES IN EXPLAINING THE LEVEL OF BIDS: ENDOGENOUS VARIABLE IS BID, ABSOLUTE t STATISTICS IN PARENTHESES, RESTRICTED TO SIGNAL VALUES $\underline{x} + \varepsilon \leq x_i \leq \bar{x} - \varepsilon$

Private-information auctions								
Regression	Private signal	Number of bidders	Cash Balance prior to bid	Auction period (normalized) ^a	Y	Cash balance × private signal	F statistic for $\varepsilon \times$ subject ^b [P]	
1	0.994 (439)	0.459 (2.50)	-0.080 (2.53)	2.50 (2.70)	-0.251 (2.02)	—	8.32 [0.001]	
2	0.994 (437)	0.531 (2.91)	-0.031 (1.19)	—	-0.242 (1.94)	—	8.58 [0.001]	
3 ^c	0.993 (428)	0.456 (2.44)	-0.085 (2.61)	2.56 (2.71)	-0.267 (2.10)	—	8.61 [0.001]	
4	0.995 (211)	0.462 (2.51)	-0.068 (1.37)	2.50 (2.70)	-0.256 (2.04)	-0.0001 (0.32)	8.20 [0.001]	
Public-information auctions								
Regression	Private signal	Number of bidders	Cash Balance prior to bid	Auction period (normalized) ^a	Public information	Private signal dummy (N = 6 or 7)	Public information dummy (N = 6 or 7)	F statistic for $\varepsilon \times$ subject ^b [P]
5	0.426 (8.94)	1.09 (1.11)	-0.151 (2.32)	7.20 (3.29)	0.564 (12.7)	-0.239 (4.40)	0.265 (5.13)	3.36 [0.001]

Note: In all cases, a fixed-effects regression model with individual-subject dummy variables was used.

^aNormalized auction-period variable is $[1/\text{last auction period}] \times [\text{auction period}]$.

^bCoefficient estimates for ε are not unique given the $\varepsilon \times$ subject dummy variables. There was a statistically significant main effect due to ε in all cases.

^cExcludes subjects who went bankrupt.

ing that each \$1 increase in cash balances produces an \$0.08 decrease in bids. The auction-period variable is positive and statistically significant, indicating that what multiperiod effects are present in the data tend to mitigate the winner's-curse effect in early auction periods, as we would expect in a design with positive expected profits.¹²

Specification 2 drops the auction-period variable. Under this specification the cash-balance variable, while still negative, is sharply reduced in size and fails to achieve statistical significance at conventional levels. This, in conjunction with specification 1,

suggests that, as the threat of being excluded on account of bankruptcy from potentially profitable auctions grew smaller, subjects bid more, but this effect was weaker for bidders with larger cash balances (average cash balances were relatively flat or decreased moderately over the initial auction periods and were increasing over later auction periods, as the survivors generally earned positive profits and bankrupt bidders were eliminated from the auction once their cash balances turned negative).

Specification 3 repeats the first specification but excludes those bidders who went bankrupt. There is virtually no difference in parameter estimates from the first specification. In excluding subjects who went bankrupt, we excluded the two subjects whose cash balances dropped below the \$4.50 threshold employed in our simulations. The fact that the coefficient for the cash-balance variable remains virtually unchanged, in conjunction with our simulation

¹²Given the varied and often puzzling regression specifications reported in HL, in conjunction with their multiperiod limited-liability argument, it is hard to believe that they did not attempt a specification including a time-trend variable. The fact that our results completely contradict their multiperiod example suggests why such a specification is not reported (Edward Leamer, 1983).

results, provides conclusive evidence that the negative coefficient associated with this variable is not a result of limited-liability forces.¹³

Specification 4 includes the cash-balance \times private-signal interaction term that HL are so fond of in their regressions. A cash-balance \times signal interaction term with the sign they report (which is negative; see specification 6 in table 3 in HL) is puzzling given their limited-liability argument, since it implies that a low cash balance will promote higher bidding with higher signal values and lower bidding with lower signal values, whereas, if anything, we would anticipate the opposite pattern in response to any limited-liability forces.¹⁴ The estimated size of the coefficient value is trivial (-0.00011) and has a t statistic of 0.33 in our specification, hardly suggestive of a robust force underlying auction behavior.

We also checked for multiperiod effects and cash-balance effects in auctions with public and private information. After all, these are also common-value auctions with the possibility for losses, so that HL's limited-liability argument should apply here as well. Regression 5 reports our preferred specification, which matches the underlying specification of the first regression in table 8 of KL, with the addition of an auction-period variable and a cash-balance variable. Here too, we observe a positive, statistically significant, time-trend coefficient and a negative, statistically significant, cash-balance coefficient, results that are qualitatively similar to but quantitatively stronger than those reported with private information only. Note

that, although there was considerable variability in profits in these auctions, on average they were only slightly lower than the RNNE prediction (KL, p. 911). Hence, the negative coefficient associated with the cash-balance variable here cannot be attributed to limited-liability forces either.

While the regression results in Table 2 are interesting, we caution the reader against placing undue reliance on them as there is some question regarding the stability of the coefficients across auction series. This manifests itself in the fact that we observe a statistically significant cash-balance \times auction-series interaction effect in the auctions with public and private information and a statistically significant auction-period \times auction-series interaction effect in both data sets. There is considerable variability among auction series in these coefficients, so the results in Table 2 reflect average tendencies for the experiment as a whole.¹⁵ There is a need to explore these multiperiod and cash-balance effects experimentally.

III. Additional Evidence

For an experimenter, the ultimate resolution of the limited-liability issue can only be obtained through additional experiments aimed at assessing the effects of exogenous variation in cash balances on bidding. Fortunately, there already exist some data relevant to the issues raised. First, in a series of common-value offer auctions in which the starting cash balance for two student groups was set at \$10, with the starting balance for the third group set at \$20, simple compar-

¹³We also looked at the data for the two subjects who were bidding with cash balances below \$4.50. Neither of them lasted very long, but both had very low balances: \$0.52 and \$0.23. Both bidders discounted their bids, relative to their signal values, *more* with these low cash balances than when they were bidding with the original balance of \$10.00.

¹⁴With higher signal values, expected profits conditional on winning the auction are increasing, at least up to the point at which the Y term is close to zero. It is only the possibility of negative profits at the Nash equilibrium that gives rise to a potential limited-liability problem in our design.

¹⁵Additional empirical analysis of the effects of cash balances on bidding is reported in Kagel et al. (1989), in this case for inexperienced bidders, for whom the likelihood of low cash balances and bankruptcy was substantially higher than for the experienced bidders studied here. The regressions reported in Kagel et al. (1989) show that higher cash balances generally promoted higher bidding, although we cannot distinguish whether this is a cash-balance effect or a possible learning effect, as subjects may have had to suffer real losses in order to learn to bid less.

isons of mean differences in profits between groups (controlling for ε) showed no significant differences between auction series (Douglas Dyer et al., 1989).

Second, Lind and Plott (1991) report results from a common-value auction experiment with two alternative procedural modifications designed to control for any limited-liability effect. Under one procedure, the common-value auction experiment in which bidders might lose money was conducted simultaneously with a second experiment in which subjects were making money. The second procedure involved sellers in a common-value auction in which the seller's loss occurs as an opportunity cost, so the possibility of bankruptcy does not exist. Lind and Plott find a winner's curse in both settings.

IV. Summary and Conclusions

We agree with HL's argument that limited liability for losses may result in overbidding and other (apparent) violations of Nash-equilibrium bidding theory on the part of perfectly rational bidders. However, our analysis and the results of Lind and Plott (1991) show that this argument is not relevant to our experimental design.

We are not surprised to find an HL type of comment. It is not the first time that some economists have employed acrobatics to avoid facing a phenomenon that does not agree with their prior beliefs. Both this reply and the work of Lind and Plott clarify the fact that financially motivated individuals can and do make judgmental errors in a

market setting, and the winner's curse is alive and well, and not the result of limited-liability effects as HL speculate.

REFERENCES

- Capen, E. C., Clapp, R. V. and Campbell, W. M., "Competitive Bidding in High-Risk Situations," *Journal of Petroleum Technology*, June 1971, 23, 641-53.
- Dyer, Douglas, Kagel, John H. and Levin, Dan, "A Comparison of Naive and Experienced Bidders in Common Value Offer Auctions: A Laboratory Analysis," *Economic Journal*, March 1989, 99, 108-15.
- Hansen, Robert G. and Lott, John R., Jr., "The Winner's Curse and Public Information in Common Value Auctions: Comment," *American Economic Review*, March 1991, 81, 347-61.
- Kagel, John H. and Levin, Dan, "The Winner's Curse and Public Information in Common Value Auctions," *American Economic Review*, December 1986, 76, 894-920.
- Kagel, John H., Levin, Dan, Battalio, Raymond and Meyer, Donald J., "First-Price Common Value Auctions: Bidder Behavior and the Winner's Curse," *Economic Inquiry*, April 1989, 27, 241-58.
- Leamer, Edward E., "Let's Take the Con Out of Econometrics," *American Economic Review*, March 1983, 73, 31-43.
- Lind, Barry and Plott, Charles R., "The Winner's Curse: Experiments with Buyers and with Sellers," *American Economic Review*, March 1991, 81, 335-46.
- Wilson, Robert, "Strategic Analysis of Auctions," mimeo, Stanford Business School, 1990.

Some Evidence on the Winner's Curse: Comment

By DAN LEVIN AND JAMES L. SMITH*

The theory of auctions has developed extensively since Robert B. Wilson's (1977) seminal paper. Due to the complexity of equilibrium strategies, however, empirical researchers have been slow to incorporate and test the most basic theoretical precepts.¹ Typical empirical studies estimate ad hoc bidding models, with no attempt to ascertain whether the implied behaviors are theoretically plausible.

The recent paper by Stuart E. Thiel (1988) attempts to bridge this gap. Thiel obtains closed-form equilibrium bidding functions that are theoretically motivated and linear in parameters, and which facilitate empirical estimation and testing. If widely applicable, Thiel's empirical approach would constitute a major methodological breakthrough.

Unfortunately, Thiel's approach applies only in special cases that are of limited practical interest. Linear bidding strategies emerge only under circumstances that are unlikely in the real world. The limited range of Thiel's approach may not be apparent to casual readers of his paper.

Even when linear strategies do exist, they are not unique under the special assumptions of Thiel's model. For each Nash equilibrium in linear strategies, there exists a

related family of nonlinear strategies. Thus, further justification must be found for basing empirical research on the linear specification.

We also note a significant error in Thiel's work. The symmetric strategies he derives on the basis of order statistics do not constitute a Nash equilibrium. We derive proper expressions for the symmetric Nash strategies and discuss a specification error in Thiel's regression analysis that would account for the mixed results obtained in his application to the highway-construction industry.

I. Linear Nash Strategies

Like Thiel, we consider a common-value auction in which each bidder observes a random signal X_i , with variance σ^2 , which is an estimate of the unknown value θ of the item being auctioned. $F(X_i|\theta)$ represents the conditional distribution of the i th bidder's estimate, and $\phi(\theta|X_i)$ represents the bidder's posterior density for θ given the signal X_i . Let $B(X)$ represent the symmetric equilibrium strategy used by each of n bidders, with $B'(\cdot) > 0$ such that the inverse $r(B) = B^{-1}$ is defined. If all other bidders are using this strategy, then the i th bidder's problem may be written as

$$\max_{B(X)} \int_{\Sigma \theta|X} (\theta - B) F(r(B)|\theta)^{n-1} \phi(\theta|X) d\theta$$

with first-order condition given by

$$0 = \int_{\Sigma \theta|X} [(\theta - B)(n-1)F(r(B)|\theta)^{n-2} \\ \times f(r(B)|\theta)r'(B) \\ - F(r(B)|\theta)^{n-1}] \phi(\theta|X) d\theta.$$

*Department of Economics, University of Houston, Houston, TX 77204-5882. The authors thank John Kagel for initially calling their attention to the issues raised in this paper and for providing some of the motivation to work them out. The authors also thank an anonymous referee who provided helpful comments on an earlier draft. The research leading to this paper was supported by the University of Houston Energy Laboratory. Dan Levin also acknowledges support from the Sloan Foundation and the National Science Foundation. These parties are not responsible for the contents of the current paper.

¹Richard Engelbrecht-Wiggans and Robert J. Weber (1979) have shown that symmetric equilibrium strategies do not even exist in closed form, except in simplified common-value auction settings.

At the symmetric equilibrium, $r(B) = X$ and $r'(B) = B'(X)^{-1}$. Thus,

$$(1) \quad B'(X)$$

$$= \frac{\int_{\Sigma \theta | X} (\theta - B)(n-1)F(X|\theta)^{n-2}f(X|\theta)\phi(\theta|X) d\theta}{\int_{\Sigma \theta | X} F(X|\theta)^{n-1}\phi(\theta|X) d\theta}.$$

A symmetric equilibrium must satisfy this differential equation and possibly other boundary conditions dictated by the context of the bidding problem.²

Thiel implicitly imposes the following three restrictions which guarantee the existence of linear strategies.

- 1) Each bidder's prior distribution of value is diffuse: $\phi(\theta)$ is constant for all θ .
- 2) Estimation errors are statistically independent of the item's true value: $F'(X_i - \theta|\theta) = f(X_i - \theta|\theta) = f(X_i - \theta)$.
- 3) Each bidder's estimate of the value is unbiased: $E(X_i) = \theta$.

Under restrictions 1–3, equation (1) reduces to

$$(2) \quad B'(X) + K_1 B(X) + K_2 - K_1 X = 0$$

where

$$K_1 = \frac{n(n-1)}{\sigma} \int_{\Sigma z} F(z)^{n-2} f(z)^2 dz > 0$$

$$K_2 = n(n-1) \int_{\Sigma z} z F(z)^{n-2} f(z)^2 dz$$

$$z = \frac{(X - \theta)}{\sigma}$$

(see Appendix). The general solution to

equation (2) is

$$(3) \quad B(X) = X - \frac{(1 + K_2)}{K_1} + \beta \exp(-K_1 X)$$

(see Appendix). However, additional conditions must be imposed to ensure individual rationality of the proposed solution. In the present case, these conditions imply $\beta \leq 0$ (see Appendix). Only in the special case of $\beta = 0$ is the bidding function linear.

Linear strategies are not unknown in the literature. Previous researchers have noted their existence under the restrictions of Thiel's model. Richard Engelbrecht-Wiggans and Robert J. Weber (1979) were among the first to discuss the necessity of a diffuse prior (restriction 1). Michael H. Rothkopf (1980) and Robert L. Winkler and Daniel G. Brooks (1980) provided examples in which independent estimation errors (restriction 2) were used to derive linear bidding functions. More recently Wilson (1988) showed that bidding strategies converge to linearity as normally distributed priors become diffuse in the limit.

The main problem is that the necessary conditions for linear strategies rule out most cases that are of any empirical interest. For example, the restriction to a diffuse prior eliminates auctions with bounds on the value of the item being sold. Zero is often a relevant bound—one that imposes structure on bidder's priors. The announcement of presale value estimates or reservation prices by the seller is another real-world phenomenon that undermines diffuse priors and linear strategies. In practice, it is hard to think of auctions in which prior information would not limit Thiel's approach, even in the highway-construction market to which his method is applied.

II. Examples

Our derivation of equilibrium strategies highlights two further problems with Thiel's (1988) paper. First, the solution in linear strategies is not a unique equilibrium. Equation (3) admits a continuous family of

²Typically, initial conditions stemming from the boundedness of value estimates identify a unique bidding function, as in Douglas K. Reece (1978) and Robert B. Wilson (1988). When value estimates are unbounded, as in Thiel's model, the standard conditions do not apply.

nonlinear strategies as well. Whether bidders would select the linear strategy is uncertain.³ Second, the bidding function Thiel derives by manipulating order statistics is not a Nash equilibrium. We present two examples that illustrate these points.

A. The Uniform Distribution

Let value estimates follow an unbiased uniform distribution over the interval $[\theta - \varepsilon, \theta + \varepsilon]$. As shown in the Appendix, any function of the following form constitutes a symmetric equilibrium:

$$B(X) = X - \varepsilon + \beta \exp[-nX/2\varepsilon]$$

for any $\beta \leq 0$.

The linear form, $B(X) = X - \varepsilon$, is obtained as a special case by choosing $\beta = 0$.

Thiel's bidding function in the uniform model reduces to

$$B(X) = X - \varepsilon(n^2 - n + 2)/(n^2 - 1)$$

which corresponds to the equilibrium only when $n = 3$.

B. The Normal Distribution

Let value estimates follow an unbiased normal distribution with mean θ and variance σ^2 . Then, the transformed value estimates $(X - \theta)/\sigma$ follow the standardized normal distribution function, $G(0,1)$. As shown in the Appendix, any function of the following form constitutes a symmetric equilibrium:

$$(4) \quad B(X) = X - \alpha_n \sigma + \beta \exp[-\xi_{n:n} X / \sigma]$$

for any $\beta \leq 0$

³Refinement of the Nash equilibrium concept will not eliminate the nonlinear strategies, since all Nash equilibria here are strict equilibria. This implies that each of the multiple equilibria is perfect in the sense of Reinhard Selten (1975). See Selten (1975) for the discrete case and Ronald M. Harstad and Levin (1986) for the continuous case.

where

$$\alpha_n = \frac{\int_{-\infty}^{+\infty} t^2 dG(t)^n}{\int_{-\infty}^{+\infty} t dG(t)^n}$$

$$= \frac{\int_{-\infty}^{+\infty} t^2 dG(t)^n}{\xi_{n:n}}$$

and where $\xi_{n:n}$ is Thiel's (1988) symbol for the expected value of the highest (standardized) value estimate received by any bidder. The linear form of equation (4), obtained by choosing $\beta = 0$, corresponds exactly to the bidding function Wilson (1988) obtains when his prior is made diffuse in the limit, and it agrees with Winkler and Brooks's (1980) result for the special case $n = 2$. For comparison with Thiel (1988), the linear version of our equation (4) can be rewritten as

$$(5) \quad B(X) = X - \sigma(1 + K_2)/\xi_{n:n}.$$

Thiel's bidding function, evaluated in the normal case, again departs from equilibrium:

$$B(X) = X - \sigma \left[\xi_{n:n} + \frac{(\pi/2)^{1/2}}{(n-1)} \right].$$

What is awry? Thiel does not properly formulate the relationship between the bidder's estimate and the probability of winning. The bidder's estimate affects his probability of winning in two ways: first, through its impact on the perceived distribution of competing bids, and second, through its impact on the level of his own bid. Thiel's differentiation of his equation 2 ignores the first effect.

III. A Warning to Empiricists

We have argued that the assumptions underlying Thiel's (1988) model are seldom satisfied in the real world. Is it possible nevertheless that linear bidding functions provide approximations that are good

enough for empirical work? The evidence is not encouraging.

Some evidence is found in Thiel's own study of the highway-construction industry. He noted the "wholesale breakdown" of the linear model in cases where the value of the contract is estimated by the state and divulged prior to bidding (p. 886). In cases where presale estimates were not divulged, Thiel's linear model seemed to fit better. What could account for this pattern? Divulging the state's estimate imposes structure on the bidders' prior distributions. If priors are not diffuse, Thiel's linear specification is not admissible (i.e., there is a specification error in the functional form of Thiel's regression equation). The cases in which presale estimates were not divulged to potential bidders do not so obviously violate Thiel's maintained hypothesis of diffuse priors. In these cases, Thiel's linear functional form is at least admissible, although the constant term in the regression equation would be misspecified, due to his mathematical error. In summary, suspected violations of Thiel's stringent paradigm tend to explain the pattern of his empirical results.⁴

If linear bidding functions are used for convenience in empirical work, how far out of equilibrium are they likely to be? This depends upon the context, so no general answer is possible. However, some illustrative results are available. In simulation studies of auctions styled after the market for offshore petroleum leases, Engelbrecht-Wiggans (1979, 1983) found that any bidder could increase expected profits up to 40 percent by departing unilaterally from an equilibrium based on second-best linear strategies and submitting optimal (nonlinear) bids instead. Bidders have strong incentives not to use linear strategies except under extremely unlikely circumstances. They have no incentive to use Thiel's linear strategies under any circumstances.

⁴Engelbrecht-Wiggans and Weber (1979) warn specifically against linearity in cases where presale value estimates are divulged by the seller.

APPENDIX

Consider a common-value auction in which each of n bidders observes a random estimate X_i of the unknown value θ of the item being auctioned. Let $F(X_i|\theta)$ represent the conditional distribution function of the i th bidder's estimate and let $f(X_i|\theta)$ represent the density. Each bidder's posterior density for θ can be written as $\phi(\theta|X_i) = f(X_i|\theta)\phi(\theta)/f(X_i)$, where $\phi(\theta)$ represents the bidder's prior distribution of values and $f(X_i)$ represents the marginal distribution of signal X_i . We let $B(X)$ represent the symmetric equilibrium strategy used by each bidder, with $B'(\cdot) > 0$, such that the inverse $r(B) = B^{-1}$ is well defined.

As shown in the text, the necessary condition for the i th bidder can be written as

$$(A1) \quad B'(X) = \frac{\int_{\Sigma \theta|X} (\theta - B)(n-1)F(X|\theta)^{n-2}f(X|\theta)\phi(\theta|X) d\theta}{\int_{\Sigma \theta|X} F(X|\theta)^{n-1}\phi(\theta|X) d\theta}$$

While this expression is not tractable for the general case, it can be evaluated under restrictions 1–3, which characterize Thiel's model. Restriction 1 (diffuse prior) implies that $\phi(\theta) = k$, a constant for all θ . Restrictions 2 and 3 (unbiased and independent errors) affect the form of $f(X_i|\theta)$. After applying restriction 1 to equation (A1) and simplifying, we have

$$(A2) \quad B'(X) = \frac{\int_{\Sigma \theta|X} (\theta - B)(n-1)F(X|\theta)^{n-2}f(X|\theta)^2 d\theta}{\int_{\Sigma \theta|X} F(X|\theta)^{n-1}f(X|\theta) d\theta}$$

For future reference, we note that

$$\begin{aligned} F(X|\theta) &= \int_{-\infty}^X f(\tau|\theta) d\tau = \int_{-\infty}^{X-\theta} f(t|\theta) dt \\ &= \int_{-\infty}^{X-\theta} f(t) dt = F(X-\theta) \end{aligned}$$

where $t = \tau - \theta$. The third equality is due to restriction 2, while the last equality is by definition. We can now rewrite equation (A2) as

$$(A3) \quad B'(X) = \frac{\int_{\Sigma \theta|X} [(X-B) + (\theta-X)](n-1)F(X-\theta)^{n-2}[\sigma f(X|\theta)]^2 \sigma^{-2} d\theta}{\int_{\Sigma \theta|X} F(X-\theta)^{n-1}[\sigma f(X|\theta)] \sigma^{-1} d\theta}$$

Let $z = (X - \theta)/\sigma$ and $dX/dz = \sigma$. Then, $f(z) = \sigma f(X|\theta)$ and $d\theta = -\sigma dz$. Applying these transforma-

tions to equation (A3) yields

$$(A4) \quad B'(X) = \frac{\int_{\Sigma_z} \left(\frac{X-B}{\sigma} - z \right) (n-1) F(z)^{n-2} f(z)^2 dz}{\int_{\Sigma_z} F(z)^{n-1} f(z) dz}.$$

The denominator is simply $1/n$, so equation (A4) can be rewritten as

$$(A5) \quad B'(X) + K_1 B(X) + K_2 - K_1 X = 0$$

where

$$K_1 = \frac{n(n-1)}{\sigma} \int_{\Sigma_z} F(z)^{n-2} f(z)^2 dz > 0$$

$$K_2 = n(n-1) \int_{\Sigma_z} z F(z)^{n-2} f(z)^2 dz.$$

$B(X) = X - (1 + K_2)/K_1$ solves equation (A5). A general solution to the homogeneous part of (A5) is $B(X) = \beta \exp(-K_1 X)$. Thus, the general solution to equation (A1) is

$$(A6) \quad B(X) = X - (1 + K_2)/K_1 + \beta \exp(-K_1 X).$$

Individual rationality restricts β to be nonpositive. To see why, assume the contrary, $\beta > 0$. Then, if $1 + K_2 > 0$, for any $X < -(1/K_1) \ln[(1 + K_2)(\beta K_1)]$, we would have $B(X) > X$, which assures negative expected profits for that X . If $1 + K_2 \leq 0$, then, for any X , we again have $B(X) > X$ and negative expected profits. This result draws on the fact that the estimate X is unbounded in Thiel's (1988) model, since he assumes unbounded θ and independent errors. We therefore conclude that $\beta \leq 0$.

We now characterize the complete family of Nash equilibria that arise in two special cases often treated in the literature.

A. Uniformly Distributed Errors

We assume that X_i is distributed uniformly over the interval $[\theta - \varepsilon, \theta + \varepsilon]$. In this case, equation (A1) reduces to

$$(A7) \quad B'(X) = \frac{\int_{X-\varepsilon}^{X+\varepsilon} (\theta - B)(n-1)(X + \varepsilon - \theta)^{n-2} d\theta}{\int_{X-\varepsilon}^{X+\varepsilon} (X + \varepsilon - \theta)^{n-1} d\theta}.$$

Integrating the numerator by parts and rearranging, we obtain

$$(A8) \quad B'(X) + \frac{nB(X)}{2\varepsilon} - \left[\frac{n(X - \varepsilon)}{2\varepsilon} + 1 \right] = 0.$$

The general solution to this first-order, linear differential equation is given by

$$(A9) \quad B(X) = X - \varepsilon + \beta \exp[-nX/2\varepsilon].$$

The individual rationality condition ($\beta \leq 0$) in this case assures $B(X) \leq X - \varepsilon$ and guarantees positive profits for the winner, since θ cannot be more than ε away from the observed X .

Evaluating Thiel's (1988) equation 3 when X_i is assumed to follow a uniform distribution over the interval $[\theta - \varepsilon, \theta + \varepsilon]$ produces the linear function

$$(A10) \quad B(X) = X - \varepsilon[(n^2 - n + 2)/(n^2 - 1)]$$

which violates our necessary condition for equilibrium.

B. Normally Distributed Errors

We assume that X_i is distributed normally with mean θ and variance σ^2 . Integrating the expressions for K_1 and K_2 by parts then gives

$$K_1 = \sigma^{-1} \int_{-\infty}^{+\infty} t dG(t)^n$$

$$1 + K_2 = \int_{-\infty}^{+\infty} t^2 dG(t)^n.$$

Thus, equation (A6) becomes

$$(A11) \quad B(X) = X - \alpha_n \sigma + \beta \exp[-\xi_{n:n} X / \sigma]$$

where

$$\alpha_n = \frac{\int_{-\infty}^{+\infty} t^2 dG(t)^n}{\int_{-\infty}^{+\infty} t dG(t)^n}$$

$$= \frac{\int_{-\infty}^{+\infty} t^2 dG(t)^n}{\xi_{n:n}}$$

and where $\xi_{n:n} = \sigma K_1$ represents the expected value of the highest (standardized) value estimate received by any bidder.

As argued for the general case, we cannot have $\beta > 0$. On the other hand, $\beta \leq 0$ implies $B(X) \leq X - \alpha_n \sigma < X - K_1 \sigma = X - \xi_{n:n}$. (The strict inequality is based on $\alpha_n = (1 + K_2)/K_1$, which by Schwarz's inequality must exceed K_1). Since the right-most term in the string of inequalities represents the expected value of θ given that X is the highest estimate, the condition $\beta \leq 0$ assures positive expected profits.

For comparison with Thiel (1988), assume $\beta = 0$ and rewrite equation (A11) as

$$(A12) \quad B(X) = X - \sigma(1 + K_2)/\xi_{n:n}.$$

Evaluating Thiel's function when X_i is assumed to follow a normal distribution gives

$$(A13) \quad B(X) = X - \sigma \left[\xi_{n:n} + \frac{(\pi/2)^{1/2}}{n-1} \right].$$

REFERENCES

- Engelbrecht-Wiggans, Richard, "Bidding in Auctions with Multiplicative Lognormal Errors: An Example," Cowles Foundation Discussion Paper No. 500R, Yale University, April 17, 1979.
- _____, "An Introduction to the Theory of Bidding for a Single Object," in Richard Engelbrecht-Wiggans, Martin Shubik, and Robert M. Stark, eds., *Auctions, Bidding, and Contracting: Uses and Theory*, New York: New York University Press, 1983, 53-103.
- _____, and Weber, Robert J., "On the Non-existence of Multiplicative Equilibrium Bidding Strategies," Cowles Foundation Discussion Paper No. 523, Yale University, April 18, 1979.
- Harstad, Ronald M. and Levin, Dan, "An Interpretation of Perfect Equilibrium for Auctions," mimeo, Department of Economics, University of Houston, August 1986.
- Reece, Douglas K., "Competitive Bidding for Offshore Petroleum Leases," *Bell Journal of Economics*, Autumn 1978, 9, 369-84.
- Rothkopf, Michael H., "On Multiplicative Bidding Strategies," *Operations Research*, May-June 1980, 28, 570-5.
- Selten, Reinhard, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 1975, 4 (1), 25-55.
- Thiel, Stuart E., "Some Evidence on the Winner's Curse," *American Economic Review*, December 1988, 78, 884-95.
- Wilson, Robert B., "A Bidding Model of Perfect Competition," *Review of Economic Studies*, October 1977, 44, 511-8.
- _____, "Strategic Analysis of Auctions," in Robert Aumann and Sergiu Hart, eds., *Handbook of Game Theory*, (draft dated October 14, 1988), Amsterdam: North-Holland, forthcoming.
- Winkler, Robert L. and Brooks, Daniel G., "Competitive Bidding with Dependent Value Estimates," *Operations Research*, May-June 1980, 28, 603-13.

Ski-Lift Pricing with Applications to Labor and Other Markets: Comment

By TYLER COWEN AND AMIHAI GLAZER*

In a recent article for this *Review*, Robert Barro and Paul Romer (1987) argue that nominal price stickiness can be compatible with the optimal provision and utilization of excludable goods. The paradigmatic case examined by Barro and Romer is the ski lift: instead of charging for each ride, a comparable equilibrium can be attained by charging a single price for access to the ski lift and using a quantity constraint to limit an individual's number of rides.

In particular, the more people show up to ski, the fewer rides each person can get, so that the real price of a single run down the slope increases. Crowding costs thus serve as a substitute for the flexibility of nominal prices by automatically changing the real prices individuals face.

Although Barro and Romer develop some imaginative applications of their analysis, they do not indicate that the basic model is a rediscovery of the theory of clubs (James Buchanan, 1965; Richard Cornes and Todd Sandler, 1986; Suzanne Scotchmer, 1985). In fact, Scotchmer (1985) explicitly mentions ski lifts as an example of clubs. Barro and Romer cite Frank Knight (1924), whose article was a forerunner of the theory of clubs, but do not mention that many of their primary results are well known in the public-economics literature.

The theory of clubs is concerned with the optimal size of a congestible facility, the

optimal intensity of its use, and pricing methods that lead to profit maximization and to social optimality. These questions correspond to the issues in Barro and Romer which involve the optimal number of skiers and the optimal price for a ski-lift ticket. Membership fees for clubs are analogous to ticket prices for the ski lift. The number of rides that skiers obtain is analogous to congestion in club theory; congestion reduces each user's utility as the number of people who visit the facility increases.

In an undergraduate text, Robin Boadway and David Wildasin (1984 p. 97) note that the optimal number of users of a club is reached when the per-person price equals the marginal congestion cost imposed by a single club member. In the context of the ski-lift example, the price of a lift ticket must equal the value of the rides that other skiers are deprived of. Romer and Barro (1987 pp. 878-9, 886) obtain an analogous result.

Many of the conclusions Barro and Romer obtain by making stringent assumptions about the effect of increased congestion (each skier obtains proportionally fewer rides) are derived by the theory of clubs from more general assumptions about the effect of congestion. Boadway and Wildasin (1984 p. 98) present the following summary of the results of club theory for the case in which intensity of use is fixed:

If exclusion is possible without cost and if the costs of providing the good are constant, the competitive market mechanism can be relied upon to provide an economy of many individuals with the correct size of facility and the correct number of members, and, therefore, the correct number of such facilities for the entire population.

*Cowen: Department of Economics, George Mason University; Glazer: Department of Economics, University of California, Irvine, CA 92717. We thank Suzanne Scotchmer, Esko Niskanen, and an anonymous referee for their comments, and members of the Transportation Economics Study Group at U.C. Irvine for stimulating discussions.

Similar conclusions are reached under yet more general assumptions by Scotchmer (1985).

Other results in Barro and Romer correspond to results in club theory: markets attain the optimum even if different ski areas vary in quality, that is, even if clubs congest at different rates. Also, the competitive solution is socially optimal when consumers have heterogeneous preferences if separate ski-lift areas serve consumers with different preferences (see Cornes and Sandler, 1986 Ch. 11).

Though there is no novelty in the problems Barro and Romer treat and no novelty in the conclusion that competitive markets can efficiently provide congestible goods, Barro and Romer do extend the theory of clubs in two important directions. First, they show that, under some conditions, efficiency can be attained by an entry fee alone; the effect of increased congestion on demand can obviate the need for imposing a price on each ride. (In contrast, Eitan Berglas [1976] and Scotchmer [1985, 1987] demonstrate that, in general, efficiency requires imposing a price for each ride.) Second, Barro and Romer offer the interesting possibility that, in some cases, both profit-maximization and efficiency imply constant prices following changes in demand.

REFERENCES

- Barro, Robert J. and Romer, Paul M., "Ski-Lift Pricing, with Applications to Labor and Other Markets," *American Economic Review*, December 1987, 77, 875-90.
- Berglas, Eitan, "On the Theory of Clubs," *American Economic Review*, May 1976, 66, 116-21.
- Boadway, Robin W. and Wildasin, David E., *Public Sector Economics*, Boston: Little, Brown, 1984.
- Buchanan, James M., "An Economic Theory of Clubs," *Economica*, February 1965, 32, 1-14.
- Cornes, Richard and Sandler, Todd, *The Theory of Externalities, Public Goods, and Club Goods*, Cambridge, U.K.: Cambridge University Press, 1986.
- Knight, Frank H., "Some Fallacies in the Interpretation of Social Cost," *Quarterly Journal of Economics*, August 1924, 38, 582-606.
- Scotchmer, Suzanne, "Two-Tier Pricing of Shared Facilities in a Free-Entry Equilibrium," *Rand Journal of Economics*, Winter 1985, 16, 456-72.
- _____, "Competitive Equilibrium and the Core in Club Economies With Anonymous Crowding" *Journal of Public Economics*, 1987, 24, 159-73.

Ski-Lift Pricing, with Applications to Labor and Other Markets: Reply

By ROBERT J. BARRO AND PAUL M. ROMER*

After we wrote our paper on pricing at ski areas (Barro and Romer, 1987), we received several comments like that of Tyler Cowen and Amihai Glazer (1991), noting that our results could be interpreted as special cases of results that have been established in club theory. We also received comments claiming that our results were clearly wrong because they contradicted standard results in club theory. (We now sympathize with the economist who was reportedly told by a critic that "your result is trivial, wrong, and I already proved it.")

When we wrote our paper, we were aware that ski areas were sometimes mentioned as examples of club goods (see, for example, the paper by Suzanne Scotchmer [1985]). We did not pursue the literature on clubs because our problem contained no club goods. The goods we dealt with, rides up a chair lift, are conventional private goods. Our result also contradicted the one result from club theory that we knew. The standard club analysis of a ski area treats it as a shared facility. Each member in the club decides how intensively to use this facility. In this setting, Eitan Berglas (1976) first showed that both a membership fee and a usage fee are necessary to support an efficient allocation. Our finding was that queues with no usage fee support an efficient allocation.

Nevertheless, Cowen and Glazer are correct that our results have much in common with results from club theory. Specifically, our results can be interpreted as special cases of results for clubs that are not subject to variable intensity of usage. There are

therefore two interesting questions: why does the analysis of a problem with private goods resemble the analysis of a problem with club goods? and why does our problem resemble a club problem in which people cannot vary how often they use the shared facilities?

The answer to both of these questions depends on the observation that it is possible to treat a bundle of private goods as a club good. Under the usual definition, both club goods and private goods are excludable. What distinguishes a club good is the assumption that it is partially nonrival. Because it is less than fully rival, many people can use a club good at the same time. Because the nonrivalry is only partial, increased usage by one person reduces the benefits of the good enjoyed by the other consumers. This reduction in benefits is typically labeled congestion.

To see that any bundle of private goods can be treated as a club good, consider a crop of apples. Individual apples are private goods, but the crop can be viewed as a good that many people consume. If one person eats an additional apple, total consumption of apples by others goes down. The loss of apple services that this person imposes on others can be treated as a congestion cost, but it is a special kind of congestion.

In the usual analysis of club goods, each individual chooses the intensity of usage taking the total intensity of usage by others as given. With some club goods, independent choices of the intensity of usage are feasible. Total capacity (for example, the size of a swimming pool) does not limit the total level of usage (the total number of person-hours spent in the pool). However, for club goods that are bundles of private goods, it does not make sense to assume that each individual can decide how much to use the club good, holding constant the

*Department of Economics, Harvard University, Cambridge, MA 02138 and Department of Economics, University of California, Berkeley, CA 94720.

usage by others. The total number of ski runs cannot exceed the total number of lift rides; the total number of apples consumed cannot exceed the supply of apples.

In most models of private goods, the supply of goods is allocated among the demanders by a price mechanism. We noted that when demanders are the same, any symmetric allocation device, including a queue, can serve this purpose. In either case (club goods or private goods), the notion of an independent decision concerning the intensity of use or amount of consumption does not make sense. In our problem, individual skiers seem to be able to choose freely how many runs to take, but in fact, each of them is constrained by the fixed supply of runs.

The point made by Cowen and Glazer can therefore be rephrased as follows: an equilibrium in which bundles of private goods are allocated by an entry fee, a zero per-unit fee, and quantity constraints is isomorphic to an equilibrium for club goods in which there is no intensity decision. This formal connection is useful, but it is important to recognize that there is something special about the implied form of congestion. Recognizing that the underlying goods are bundles of private goods, rather than true club goods, is crucial to the demonstration that queues can have negligible efficiency costs. Because the goods are private, the associated congestion costs take on the same form imposed by the person who eats an extra apple. This congestion reflects the fact that supply is fixed. There cannot be too much congestion.

We started with bundles of private goods and have now learned that they can be treated as club goods. There is a complementary analysis that starts with a general description of club goods and shows that these goods can sometimes behave like private goods. Consider a situation in which there is no decision about the intensity of use. Let the club benefits received by any member be written in the form $f(x, n)$, where x is a measure of capacity and n is the total number of users. Berglas and David Pines (1981) show that if $f(\cdot)$ is homogeneous of degree zero in x and n , then the

analysis reduces to the case of goods that are essentially private. This would be the case that applies to the apple harvest. If there are twice as many apples and twice as many identical apple-eaters, benefits received by any person remain the same.

In the third section of our paper, we reconsidered the two-roads problem suggested by Pigou and showed that, under certain functional-form assumptions, spots on the road for cars seemed to be like private goods. In the textbook analysis, free access to two different roads leads to wasteful congestion, because users will equate the average benefits on the two roads: $f(x_1, n_1) = f(x_2, n_2)$. The condition for optimality is that the marginal effect on benefits should be equated: $(d/dn_1)[n_1 f(x_1, n_1)] = (d/dn_2)[n_2 f(x_2, n_2)]$. In our analysis, we noted that the first equality implies the second if $f(x, n)$ is homogeneous of degree zero. This is precisely the point made by Berglas and Pines. If $f(\cdot)$ is homogeneous of degree zero, the problem is really one of private goods, not club goods.

In our paper, we also noted that the first equality implies the second if $f(\cdot)$ takes the more general form $f(x, n) = xn^{-\alpha}$ for any α . In private communication, David Pines observed that, subject to a nonlinear change $\sigma(\cdot)$ in the units used to measure the scale of capacity x , this more general form could be interpreted as a special case of the previous homogeneity result. The general assumption on f that causes the first equality to be equivalent to the second is that the function takes the form $f(x, n) = g(\sigma(x)/n)$. Our special case follows when $g(z) = z^{-\alpha}$ and $\sigma(x) = x^\alpha$.

While it is formally possible to treat a bundle of private goods as a club good, it may sometimes obscure, rather than clarify, the analysis. For example, in their textbook, Richard Cornes and Todd Sandler (1986 p. 190) describe a golf course at a country club as a club good that should carry a usage fee. They observe that usage fees are almost never charged and conclude that large transactions costs must be present. This assertion seems as implausible to us as the assertion that it is prohibitively expensive to charge for lift rides by the ride.

The usual club-theory prediction that zero usage fees will lead to wasteful congestion would be right if an unlimited number of golfers could freely enter a course and start hitting golf balls, but this is not how golf courses operate. The opportunity to start a round of golf at a specific time is just like the chance to ride on a particular chair on a chair-lift. It is a rival good. The same kind of analysis as for the ski area suggests that unobserved transactions costs need not be present and that the hypothesized welfare losses from mispricing and congestion may be negligible. No efficiency losses will be present if the members of the country club have similar tastes and the club imposes quantity constraints in a way that does not waste resources, for example: by a reservation system that is random and does not reward effort. A typical such reservation system, widely used in private clubs and university athletic organizations, lets individuals make a single reservation by phone, one day in advance, and only after a certain time (e.g., 8:00 A.M.).

REFERENCES

- Barro, Robert J. and Romer, Paul M., "Ski-Lift Pricing, with Applications to Labor and Other Markets," *American Economic Review*, December 1987, 77, 875-90.
- Berglas, Eitan, "On the Theory of Clubs," *American Economic Review*, March 1976, 66, 116-21.
- _____ and Pines, David, "Clubs, Public Goods, and Transportation Models," *Journal of Public Economics*, April 1981, 15, 141-62.
- Cornes, Richard and Sandler, Todd, *The Theory of Externalities, Public Goods, and Club Goods*. Cambridge, U.K.: Cambridge University Press, 1986.
- Cowen, Tyler and Glazer, Amihai, "Ski-Lift Pricing with Applications to Labor and Other Markets: Comment," *American Economic Review*, March 1991, 81, 376-7.
- Scotchmer, Suzanne, "Two-Tier Pricing of Shared Facilities in a Free-Entry Equilibrium," *Rand Journal of Economics*, Winter 1985, 16, 456-72.

A Model of Housing Tenure Choice: Comment

By YUMING FU*

In their paper in this *Review*, Vernon Henderson and Yannis Ioannides (1983) employed a two-period housing-consumption-investment model to study tenure-choice behavior. One of their major findings is that the demand for housing investment is independent of wealth, given the income path, but increases as income tilts forward. Since the demand for housing consumption increases with wealth, wealthier people are more likely to rent the housing they consume, given the income path. The purpose of this note is to show that their comparative-statics result for housing investment demand (equation 17a) contains an incorrect sign.¹ The correct result indicates that the demand for housing investment responds positively to the change in wealth, rather than to the change in the income path. As a result, whether or not wealthier people are more likely to rent the housing they consume depends on the relative magnitudes of the income elasticities of housing consumption versus investment demand. Also, Henderson and Ioannides' comparative-statics result for the savings demand (equation 17c) is incorrect. Since the demand for housing consumption, for housing investment, and for savings cannot be separated, this note corrects that error as well. The results presented in this note should make the intuition underlying Henderson and Ioannides' tenure-choice model clearer.

I. The Correct Comparative Statics and the Interpretation

Adopting the notation of Henderson and Ioannides (1983) and following their formu-

*Faculty of Commerce and Business Administration, University of British Columbia, 2053 Main Mall, Vancouver, BC, Canada V6T 1Y8. I thank Lawrence D. Jones and Robert W. Helsley for helpful comments. I also appreciate the comments of J. Vernon Henderson and Yannis M. Ioannides.

¹The incorrect sign has also been discovered by Clive Southey et al. (1990).

lation, the comparative-statics results for optimal housing consumption, \tilde{h}_c , housing investment, \tilde{h}_I , and savings, \tilde{S} , can be obtained by solving the following equation:²

$$(1) \quad \mathbf{D} \begin{bmatrix} d\tilde{h}_c \\ d\tilde{h}_I \\ d\tilde{S} \end{bmatrix} = \frac{\mathbf{b}_1}{2} \left(dy_1 - \frac{dy_2}{1+r} \right) + \frac{\mathbf{b}_2}{2} \left(dy_1 + \frac{dy_2}{1+r} \right)$$

where \mathbf{D} (the matrix given at the top of p. 382) is the Hessian obtained in differentiating the first-order conditions for \tilde{h}_c , \tilde{h}_I , and \tilde{S} , and

$$\mathbf{b}_1 = \begin{bmatrix} U_{11}R - E[V'']\tau(1+r) \\ U_{11}P_1 + E[V''P_2](1+r) \\ U_{11} + E[V''](1+r)^2 \end{bmatrix}$$

$$\mathbf{b}_2 = \begin{bmatrix} U_{11}R + E[V'']\tau(1+r) \\ U_{11}P_1 - E[V''P_2](1+r) \\ U_{11} - E[V''](1+r)^2 \end{bmatrix}$$

By using Cramer's rule and noticing that \mathbf{b}_1 is identical to the third column of \mathbf{D} , one knows immediately that both $d\tilde{h}_c$ and $d\tilde{h}_I$ are independent of the change in the income path, $dy_1 - dy_2/(1+r)$, but depend on the change in wealth, $dy_1 + dy_2/(1+r)$. However, $d\tilde{S}$ depends on the changes in both wealth and the path of income. The

²Where $P_1 = P - L - R$, and $P_2 = P(1+\theta) - L(1+r) - [T(\bar{u}) - \tau(\bar{u})]$. For comparison with the notation in the original paper, $P_1 = \xi/(1+r)$, and $P_2 = \beta + \gamma$.

$$\mathbf{D} = \begin{bmatrix} U_{11}R^2 + U_{22}f^2 + E[V'']\tau^2 & U_{11}RP_1 - E[V''P_2]\tau & U_{11}R - E[V'']\tau(1+r) \\ U_{11}RP_1 - E[V''P_2]\tau & U_{11}P_1^2 + E[V''P_2^2] & U_{11}P_1 + E[V''P_2](1+r) \\ U_{11}R - E[V'']\tau(1+r) & U_{11}P_1 + E[V''P_2](1+r) & U_{11} + E[V'']\tau(1+r)^2 \end{bmatrix}$$

solution to equation (1) yields:³

$$(2a) \quad d\tilde{h}_c = \frac{U_{11}}{D} \left\{ E[V''] E[V''(P_2 - (1+r)P_1)^2] - E^2[V''(P_2 - (1+r)P_1)] \right\} \times [R(1+r) + \tau](1+r) \left(dy_1 + \frac{dy_2}{1+r} \right)$$

$$(2b) \quad d\tilde{h}_I = -\frac{U_{11}U_{22}}{D} f^2 E[V''(P_2 - (1+r)P_1)] \times (1+r) \left(dy_1 + \frac{dy_2}{1+r} \right)$$

$$(2c) \quad d\tilde{S} = \frac{1}{2} \left(dy_1 - \frac{dy_2}{1+r} \right) + \frac{1}{2D_{-I}} \times \left\{ U_{11}U_{22}f^2 - U_{11}E[V''] \times [R^2(1+r)^2 - \tau^2] - U_{22}f^2E[V'']\tau(1+r)^2 \right\} \times \left(dy_1 + \frac{dy_2}{1+r} \right) - \left\{ P_1 + \frac{1}{D_{-I}} (U_{11}[R(1+r) + \tau]R + U_{22}f^2(1+r)) \right\} \times E[V''(P_2 - (1+r)P_1)] \Big\} d\tilde{h}_I$$

where

$$D = U_{11}U_{22}f^2E[V''(P_2 - (1+r)P_1)^2] + \{U_{11}[R(1+r) + \tau]^2 + U_{22}(1+r)^2f^2\} \times \{E[V'']E[V''(P_2 - (1+r)P_1)^2] - E^2[V''(P_2 - (1+r)P_1)]\}$$

is the determinant of \mathbf{D} and where

$$D_{-I} = U_{11}U_{22}f^2 + U_{11}E[V''] [R(1+r) + \tau]^2 + U_{22}f^2E[V'']\tau(1+r)^2$$

is the determinant of a reduced \mathbf{D} (with its second row and second column removed) and is positive.

The correct comparative-statics result for \tilde{S} , shown by equation (2c), indicates that savings will change to offset the change in the income path and that, given the income path, it will also change directly with wealth, since the income elasticities of consumption in the two periods are, in general, not equal. Furthermore, \tilde{S} has to decrease to meet the increased demand for housing investment as wealth increases. The positive relationship between housing investment demand and wealth is indicated by equation (2b). Let $A = -V''/V'$ be the coefficient of absolute risk aversion. Using a first-order condition,

$$(3) \quad E[V'(P_2 - (1+r)P_1)] = 0$$

and substituting AV' for $-V''$ in (2b), it can be shown that $d\tilde{h}_I/d(y_1 + y_2/(1+r))$ is positive if $dA/dW < 0$ and that it is 0 for $dA/dW = 0$, where W is the second-period

³It can be shown that (2a) is equivalent to equation 17b in Henderson and Ioannides (1983).

wealth.⁴ In other words, the housing investment demand increases with wealth if the coefficient of absolute risk aversion is decreasing. The investment demand remains unchanged if wealth is unchanged or if the coefficient of absolute risk aversion is constant.

Further insights into the determinants of housing investment demand may be obtained from equation (3). Using a linear approximation for V' at $\bar{\theta} = E[\theta]$, one obtains the following:⁵

$$(4) \quad P\bar{h}_1 \approx \frac{E[P_2 - (1+r)P_1]}{PA \text{Var}(\theta)}.$$

Equation (4) shows that the value of housing investment is determined by three factors: the expected rate of return on housing investment ($E[P_2 - (1+r)P_1]/P$), the coefficient of absolute risk aversion, and the investment risk ($\text{Var}(\theta) = \text{Var}\{E[P_2 - (1+r)P_1]/P\}$). In other words, investment housing in this model plays a role, not as a savings vehicle, but as a risky asset which offers an opportunity for people to trade risk-bearing for a higher return. Housing investment demand is affected by wealth only because wealth affects people's willingness to bear risks. The path of income affects neither housing investment demand nor housing consumption demand, because a perfect financial market, in which personal saving and borrowing are unrestricted,

is assumed in the model. With restrictions on personal saving and borrowing, both consumption and investment demands could be affected by the income path.

II. Conclusions

The finding in Henderson and Ioannides (1983) is counterintuitive and unsupported empirically. Indeed, Henderson and Ioannides (1987) find that the likelihood of owning a home is a positive function of lifetime wealth. They attribute the discrepancy between theoretical and observed patterns of housing tenure choice to such factors as taxes, capital market imperfections, and rental externalities. My results show that both housing consumption and investment demands may increase with wealth and, thus, that it is not necessary to rely upon those institutional and externality factors to produce a positive wealth effect on owning a home. Whether or not wealthier people are more likely to own the housing they consume depends on the magnitude of the income elasticity of housing consumption demand relative to that of housing investment demand, which in turn depends on the rate at which risk aversion decreases with wealth. If the income elasticity of investment demand exceeds that of consumption demand, then wealthier people are more likely to be owner-occupiers.

REFERENCES

- Henderson, J. Vernon and Ioannides, Yannis M., "A Model of Housing Tenure Choice," *American Economic Review*, March 1983, 53, 98-113.
- _____ and _____, "Owner Occupancy: Investment vs. Consumption Demand," *Journal of Urban Economics*, March 1987, 21, 228-41.
- Southey, Clive, Ho, Ho-Cheng and Steel, Marion, "Housing Tenure and Investment: A Re-appraisal," unpublished manuscript, University of Guelph, 1990.

⁴See appendix in Henderson and Ioannides (1983) for a proof.

⁵The linear approximation for V' at $\bar{\theta}$ is $V' \approx \bar{V}' + \bar{V}'' P\bar{h}_1(\theta - \bar{\theta})$, where \bar{V}' and \bar{V}'' are the values of V' and V'' at $\theta = \bar{\theta}$, respectively. Substituting the linear approximation for V' in equation (3), one obtains

$$\begin{aligned} 0 &= E[V'(P_2 - (1+r)P_1)] \\ &= E[P_2 - (1+r)P_1]E[V'] + PE[V'\theta] \\ &\approx E[P_2 - (1+r)P_1]\bar{V}' + \bar{V}'' P^2 \bar{h}_1 \text{Var}(\theta) \end{aligned}$$

which leads to equation (4).

ERRATUM

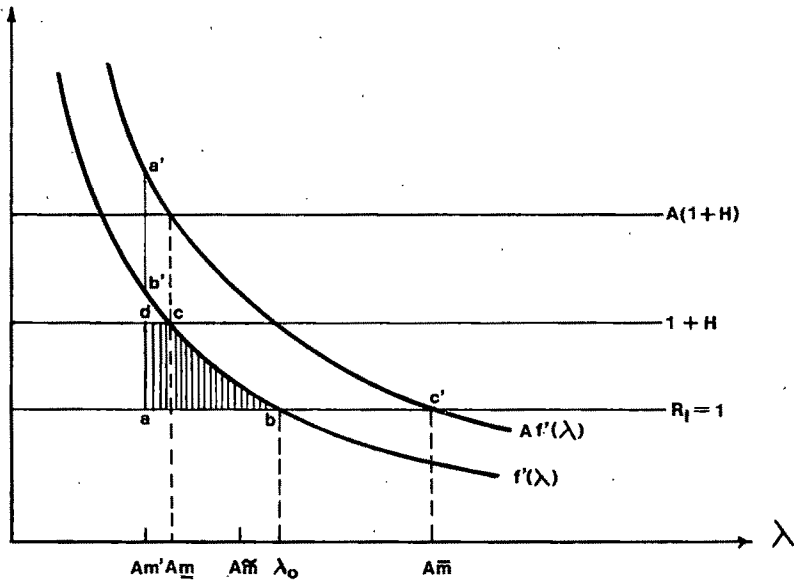
Cooperation, Harassment, and Involuntary Unemployment

By ERNST FEHR*

In my comment published in this *Review* (Volume 80, No. 3, June 1990, pp. 624-30),

*Department of Economics, University of Technology Vienna, Argentinierstrasse 8/175, A-1040 Vienna, Austria.

parts of my argument rely on Figure 1. Unfortunately, all the symbols are misspelled in this figure, and therefore my argument may have been incomprehensible. Here Figure 1 is printed with its denotations.



scenario 3 scenario 2
($0 < m \leq \underline{m}$) ($\underline{m} < m < \bar{m}$)

$\lambda_3 = A \underline{m}$

$\lambda_2 = A m$

scenario 1
($\bar{m} \leq m$)

$\lambda_1 = A \bar{m}$

FIGURE 1.

The American Economic Review

PAPERS AND PROCEEDINGS

OF THE

Hundred and Third Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

Washington, D.C., December 28-30, 1990

Program Arranged by Thomas C. Schelling

Papers and Proceedings Edited by Ronald L. Oaxaca and Wilma St. John

MAY 1991

THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

Officers

President

THOMAS C. SCHELLING
University of Maryland

President-elect

WILLIAM VICKREY
Columbia University

Vice-Presidents

HENRY J. AARON
The Brookings Institution
CLAUDIA D. GOLDIN
Harvard University

Secretary-Treasurer

C. ELTON HINSHAW
Vanderbilt University

Editor of The American Economic Review

ORLEY C. ASHENFELTER
Princeton University

Editor of The Journal of Economic Literature

JOHN PENCAVEL
Stanford University

Editor of The Journal of Economic Perspectives

JOSEPH E. STIGLITZ
Stanford University

Executive Committee

Elected Members of the Executive Committee

STANLEY FISCHER
Massachusetts Institute of Technology
LAWRENCE H. SUMMERS
The World Bank
GREGORY C. CHOW
Princeton University
SUSAN ROSE-ACKERMAN
Yale University
MICHAEL J. PIORE
Massachusetts Institute of Technology
GAVIN WRIGHT
Stanford University

EX OFFICIO Member

GERARD DEBREU
University of California-Berkeley

•Printed at Banta Company, Menasha, Wisconsin.

•Copyright © American Economic Association 1991. All rights reserved.

•No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

Correspondence relating to advertising, business matters, permissions to quote, subscriptions, and changes of address, should be sent to the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Change of address notice must be received at least six (6) weeks prior to the publication month. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

THE AMERICAN ECONOMIC REVIEW (ISSN 0002-8282), May 1991, Vol. 81, No. 2, is published five times a year (March, May, June, September, December) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Annual fees for regular membership, of which 30 percent is for a year's subscription to this journal, are: \$44.00, \$52.80, or \$61.60 depending on income. A membership also includes the *Journal of Economic Literature* and the *Journal of Economic Perspectives*. In countries other than the U.S.A., add \$16.00 for extra postage. Second-class postage paid at Nashville, TN and at additional mailing offices. POSTMASTER: Send address changes to the *American Economic Review*, 2014 Broadway, Suite 305, Nashville, TN 37203.

THE AMERICAN ECONOMIC REVIEW

VOL. 81 NO. 2

MAY 1991

PAPERS AND PROCEEDINGS

OF THE

Hundred and Third Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

Washington, D.C.

December 28–30, 1990

Program Arranged by Thomas C. Schelling

Papers and Proceedings Edited by Ronald L. Oaxaca and Wilma St. John

Copyright © AMERICAN ECONOMIC ASSOCIATION, 1991

CONTENTS

Editors' Introduction	<i>Ronald L. Oaxaca and Wilma St. John</i>	vii
Foreword	<i>Thomas C. Schelling</i>	viii

PAPERS

Richard T. Ely Lecture		
Procrastination and Obedience	<i>George A. Akerlof</i>	1
Teaching College Economics		
The Economics Major: Can and Should We Do Better than a B—?	<i>John J. Siegfried, Robin L. Bartlet, W. Lee Hansen, Allen C. Kelley, Donald N. McCloskey, and Thomas H. Tietenberg</i>	20
An Agenda for Research on Economic Education in Colleges and Universities	<i>William Becker, Robert Highsmith, Peter Kennedy, and William Walstad</i>	26
The Third Edition of the Test of Understanding College Economics	<i>Phillip Saunders</i>	32
Economics in Space		
Providing Earth Observation Data from Space: Economics and Institutions	<i>Molly K. Macauley and Michael T. Toman</i>	38
Trading Orbit Spectrum Assignments in the Space Satellite Industry	<i>Harvy J. Levin</i>	42
Torts and Orbits: The Allocation of the Costs of Accidents Involving Spacecraft	<i>Ann M. Butler and Neil A. Doherty</i>	46
The National Aerospace Plane: An American Technological Long Shot, Japanese Style	<i>Linda R. Cohen, Susan A. Edelman, and Roger G. Noll</i>	50
Tort Law as a Regulatory System		
Regulation and the Law of Torts	<i>Susan Rose-Ackerman</i>	54
The Safety and Innovation Effects of U.S. Liability Law: The Evidence	<i>Robert E. Litan</i>	59
Mispriced Equity: Regulated Rates for Auto Insurance in Massachusetts	<i>B. Glenn Blackmon, Jr. and Richard Zeckhauser</i>	65
Path Dependence in Economics: The Invisible Hand in the Grip of the Past		
Multiple Equilibria and Persistence in Aggregate Fluctuations	<i>Steven N. Durlauf</i>	70
Identifying the Hand of Past: Distinguishing State Dependence from Heterogeneity	<i>James J. Heckman</i>	75
History and Industry Location: The Case of the Manufacturing Belt	<i>Paul Krugman</i>	80
Complementarities, Momentum, and the Evolution of Modern Manufacturing	<i>Paul Milgrom, Yingyi Qian, and John Roberts</i>	84
Price Stickiness in Theory and Practice		
Why Are Prices Sticky? Preliminary Results from an Interview Study	<i>Alan S. Blinder</i>	89
Discussion	<i>Robert J. Shiller</i>	97
.....	<i>Robert J. Gordon</i>	98
.....	<i>Herschel Grossman</i>	99

Patterns of Faculty Retirement

- Projecting Faculty Retirement: Factors Influencing Individual Decisions *G. Gregory Lozier and Michael J. Dooris* 101
- Ending Mandatory Retirement in the Arts and Sciences *Sharon P. Smith* 106
- The Effects of Pensions and Retirement Policies on Retirement in Higher Education *Alan L. Gustman and Thomas L. Steinmier* 111

Economics and Conflict

- Conflicts and Attitudes Toward Risk *Stergios Skaperdas* 116
- The East European Revolution of 1989: Is It Surprising that We Were Surprised? *Timur Kuran* 121
- Nations and States: Mergers and Acquisitions; Dissolutions and Divorce ... *Donald Wittman* 126
- The Technology of Conflict as an Economic Activity *Jack Hirshliefer* 130

Greenhouse Warming

- International Trade in Carbon Emission Rights: A Decomposition Procedure *Alan S. Manne and Richard G. Richels* 135
- Towards a Comprehensive Approach to Global Climate Change Mitigation *Richard D. Morgenstern* 140
- A Sketch of the Economics of the Greenhouse Effect *William D. Nordhaus* 146

Gender and Productivity

- The Role of Off-the-Job vs. On-the-Job Training for the Mobility of Women Workers *Lisa M. Lynch* 151
- The Impact of Nonmarket Work on Market Wages *Joni Hersch* 157
- Gender Differences in Labor Market Effects of Alcoholism *John Mullahy and Jody L. Sindelar* 161

Western Europe, Eastern Europe, and the World Economy

- Europe Post-1992: Internal and External Liberalization ... *Alexis Jacquemin and André Sapir* 166
- The Challenges of German Unification for EC Policymaking and Performance *Robert F. Owen* 171
- Integration of Eastern Europe into the World Trading System *Helen B. Junz* 176
- Industry Restructuring in East-Central Europe: The Challenge and the Role for Foreign Investment *Catherine L. Mann* 181

Economic Developments and Prospects in Czechoslovakia, Yugoslavia, and Germany

- Czechoslovakia: Recent Economic Developments and Prospects... *Karel Dyba and Jan Svejnar* 185
- Economic Development in Yugoslavia in 1990 and Prospects for the Future *Janez Prasnikar and Zivko Pregl* 191
- The Economic Integration of Post-Wall Germany *Irwin L. Collier, Jr. and Horst Siebert* 196

Chinese Economic Reforms, 1979-89: Lessons for the Future

- Chinese Enterprise Behavior Under the Reforms *Roger H. Gordon and Wei Li* 202
- Why Has Economic Reform Led to Inflation? *Barry Naughton* 207
- Economic Reform of the Distribution Sector in China .. *Richard H. Holton and Terry Sicular* 212

Behavioral Finance

- Shareholder Heterogeneity: Evidence and Implications *Laurie Simon Bagwell* 218
- Investor Diversification and International Equity Markets *Kenneth R. French and James M. Poterba* 222
- Window Dressing By Pension Fund Managers *Josef Lakonishok, Andrei Shleifer, Richard Thaler, and Robert Vishny* 227
- The Rationality Struggle: Illustrations from Financial Markets *Jayendu Patel, Richard Zeckhauser, and Darryll Hendricks* 232

Economics of Drugs

Rational Addiction and the Effect of Price on Consumption	<i>Gary S. Becker, Michael Grossman, and Kevin M. Murphy</i>	237
Alcohol Consumption During Prohibition	<i>Jeffrey A. Miron and Jeffrey Zwiebel</i>	242
Who Uses Illegal Drugs?	<i>Robin Sickles and Paul Taubman</i>	248

Market Structure and the Emergence of New Technologies

R&D Competition for Product Innovation: An Endless Race	<i>Reiko Aoki</i>	252
Choosing R&D Projects: An Informational Approach	<i>Beth Allen</i>	257
The Determinants of Investment in New Technology	<i>Sarah J. Lane</i>	262
Diversification by Regulated Monopolies and Incentives for Cost-Reducing R&D	<i>Karen Palmer</i>	266

Diffusion of Development

Diffusion of Development: Post-World War II Convergence Among Advanced Industrial Nations	<i>Richard R. Nelson</i>	271
Diffusion of Development: The Soviet Union	<i>Marshall I. Goldman</i>	276
Diffusion of Development: The Late-Industrializing Model and Greater Asia	<i>Alice H. Amsden</i>	282

The Economic Impact of Immigration

Immigrants in the U.S. Labor Market: 1940-80	<i>George J. Borjas</i>	287
Immigration and Wages: Evidence from the 1980's	<i>Kristin F. Butcher and David Card</i>	292
Immigrants in the American Labor Market: Quality, Assimilation, and Distributional Effects	<i>Robert J. LaLonde and Robert H. Topel</i>	297

The History of African-American Economic Thought and Policy

W. E. B. Du Bois and the Historical School of Economics	<i>Thomas D. Boston</i>	303
Missed Opportunity: Sadie Tanner Mossell Alexander and the Economics Profession	<i>Julianne Malveaux</i>	307
The Rise and Fall of Negro Economics: The Economic Thought of George Edmund Haynes	<i>James B. Stewart</i>	311
Celestial Mechanics and the Location Theory of William H. Dean, Jr., 1930-52	<i>Julian Ellison</i>	315

Post-Communist Economic Transformation: Hungary vs. Poland

Institutional Legacies and the Economic, Social, and Political Environment for Transition in Hungary and Poland	<i>Keith Crane</i>	318
Transformation Programs: Content and Sequencing	<i>Farid Dhanji</i>	323
Foreign Economic Liberalization in Hungary and Poland	<i>Paul Marer</i>	329

Intertemporal Choice

Derivation of "Rational" Economic Behavior from Hyperbolic Discount Curves	<i>George Ainslie</i>	334
Economic Analysis and the Psychology of Utility: Applications to Compensation Policy	<i>Daniel Kahneman and Richard Thaler</i>	341
Negative Time Preference	<i>George Loewenstein and Drazen Prelec</i>	347

Learning and Adaptive Economic Behavior

Designing Economic Agents that Act Like Human Agents: A Behavioral Approach to Bounded Rationality	<i>W. Brian Arthur</i>	353
Experiments on Stable Suboptimality in Individual Behavior	<i>R. J. Herrnstein</i>	360
Artificial Adaptive Agents in Economic Theory	<i>John H. Holland and John H. Miller</i>	365

PROCEEDINGS

Minutes of the Annual Meeting	373
-------------------------------------	-----

Minutes of the Executive Committee Meetings	378
---	-----

Reports

Secretary	<i>C. Elton Hinshaw</i>	385
Treasurer	<i>C. Elton Hinshaw</i>	389
Finance Committee	<i>C. Elton Hinshaw</i>	390
Editor, <i>American Economic Review</i>	<i>Orley Ashenfelter</i>	391
Editor, <i>Journal of Economic Literature</i>	<i>John Pencavel</i>	399
Editor, <i>Journal of Economic Perspectives</i>	<i>Joseph Stiglitz</i>	401
Director, <i>Job Openings for Economists</i>	<i>C. Elton Hinshaw</i>	403
Committee on Economic Education	<i>John J. Siegfried</i>	405
Committee on the Status of Minority Groups in the Economics Profession	<i>Margaret C. Simms</i>	407
Committee on the Status of Women in the Economics Profession	<i>Nancy M. Gordon</i>	409
Committee on U.S.-China Exchanges in Economics	<i>Gregory C. Chow</i>	413
Representative to the National Bureau of Economic Research	<i>David Kendrick</i>	414
Representative to the American Association for the Advancement of Science	<i>Adam Rose</i>	416
Policy and Advisory Board of the Economics Institute	<i>Edwin S. Mills</i>	418

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. People of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

Editors' Introduction

This volume contains the *Papers and Proceedings* of the one hundred and third annual meeting of the American Economic Association. The *Proceedings* record the business activities of the Association in 1990; the annual membership meetings; the March and December meetings of the Association's officers and committees. The *Papers* constitute the greater part of the volume. They comprise contributions that fill roughly the same number of pages as two regular issues of the *American Economic Review*. We would like to take this opportunity to answer a number of commonly asked questions about the *Papers*.

Who chooses the authors? About a year in advance, the Association's President-elect, acting as program chairman, decides on the theme of the sessions and topics for which sessions will be organized. This is done after consultation and comment, both volunteered and solicited, from a wide range of individuals. (A *Call for Papers* was published annually in the Notes section of the *AER*, and now appears in the Fall issue of the *Journal of Economic Perspectives*.) The President-elect invites persons to organize these sessions. Each session organizer in turn invites several persons (usually two or three) to give papers on the theme of the session, and asks others to give comments on the papers. The program chairman decides at the time of organization which sessions will be included in this volume. Space limitations restrict the number of printed sessions. This year we are printing 22 sessions, although a total of 132 sessions were sponsored, either solely by the American Economic Association or jointly with other allied societies.

Are discussants' comments published? There has been no standard practice with regard to the publication of comments and discussions in the past. This year the President-elect decided to publish one set of

comments for one panel session. For the other sessions, names and affiliations of commentators are printed at the start of each session, permitting readers especially interested in particular comments to write to the commentator for a copy of the discussion.

What standards must the papers meet?

The guidelines under which papers are published in the *Papers and Proceedings* differ considerably from those governing regular issues of the *Review*. First, the length of papers is strictly controlled. Except in unusual circumstances they must be no more than twelve typescript pages in three-paper sessions, and eighteen typescript pages in two-paper sessions. Second, papers are not subjected to a formal refereeing process. However, a paper can be rejected if, after reading it, we conclude it is utterly without merit. This year we are pleased to report that no paper has been rejected on this ground. Third, their content and range of subject matter reflect the wishes of the President-elect to investigate and expose the current state of economic research and thinking. In most cases they are therefore exploratory and discursive, rather than formal presentations of original research.

In order to produce this volume by May, strict deadlines must be met and there is no time for communication with every author about editing changes made in order to improve content and style, and to satisfy space restrictions. Every effort is made to notify an author prior to the deadline if the paper is too long, or does not satisfy other specifications.

This year, authors for the most part were quite cooperative, and for this we are grateful.

RONALD L. OAXACA
WILMA ST. JOHN

Foreword

I could not have done my part in arranging the program for the 1990 annual meeting of the American Economic Association without the generous and proficient help of a number of people.

The staff of the American Economic Association and the Allied Social Science Associations in Nashville, responsible for negotiating with hotels and airlines, scheduling and fitting nearly 500 sessions into appropriate rooms with appropriate equipment, arranging dozens of functions to be catered, answering inquiries and publishing a faultless and informative catalogue, performed in a manner that I can only describe as heroic.

The editorial staff of the *AER Papers and Proceedings* has been helpful and patient as well as skillful. I offer them my thanks, and thanks also to a Program Committee that refereed approximately 400 submitted abstracts, selecting nearly a hundred and organizing them, with chairs and discussants, into proper sessions. They are: Henry J. Aaron, Robert Z. Aliber, Elizabeth E. Bailey, Joel Darmstadter, Paul A. David, Robert E. Dorfman, Richard Freeman, Lawrence B. Krause, Mancur Olson, Dwight H. Perkins, William Poole, Isabel V. Sawhill, David A. Wise, and Richard J. Zeckhauser.

THOMAS C. SCHELLING

RICHARD T. ELY LECTURE

Procrastination and Obedience

By GEORGE A. AKERLOF*

In this lecture I shall focus on situations involving *repeated* decisions with time inconsistent behavior. Although each choice may be close to maximizing and therefore result in only small losses, the cumulative effect of a series of repeated errors may be quite large. Thus, in my examples, decision makers are quite close to the intelligent, well-informed individuals usually assumed in economic analysis, but cumulatively they make seriously wrong decisions that do not occur in standard textbook economics.

This lecture discusses and illustrates several "pathological" modes of individual and group behavior: procrastination in decision making, undue obedience to authority, membership of seemingly normal individuals in deviant cult groups, and escalation of commitment to courses of action that are clearly unwise. In each case, individuals choose a series of current actions without fully appreciating how those actions will affect future perceptions and behavior. The standard assumption of rational, forward-looking, utility maximizing is violated. The nonindependence of errors in decision making in the series of decisions can be explained with the concept from cognitive psychology of undue salience or vividness. For example, present benefits and costs may

have undue salience relative to future costs and benefits.

Procrastination occurs when present costs are unduly salient in comparison with future costs, leading individuals to postpone tasks until tomorrow without foreseeing that when tomorrow comes, the required action will be delayed yet again. Irrational obedience to authority or escalation of commitment occurs when the salience of an action today depends upon its deviation from previous actions. When individuals have some disutility for disobedience and a leader chooses the step sizes appropriately, individuals can be induced to escalate their actions to extraordinary levels; the social psychologist Stanley Milgram (1975) led subjects to administer high levels of electrical shock to others in fictitious learning experiments. The subjects were induced into actions that were contrary to their true moral values. In the latter half of the lecture I will give examples to illustrate how sequences of errors, each error small at the time of the decision, cumulate into serious mistakes; these decisions also illustrate how laboratory conditions of isolation, carefully engineered in the Milgram experiment and necessary for the type of behavior he induced, in fact commonly occur in nonexperimental situations. Thus the sequences of errors that are the subject of this lecture are not rare and unusual, only obtainable in the laboratory of the social psychologist, but instead are common causes of social and economic pathology.

Although an analysis of behavioral pathology might initially appear to be outside the appropriate scope of economics, I shall argue that, in important instances, such pathology affects the performance of individuals and institutions in the economic and social domain. Examples include the poverty

*Department of Economics, University of California, Berkeley, CA 94720. I thank Glenn Carroll, Benjamin Hermalin, Daniel Kahneman, David Levine, Andreu Mas Colell, Charles O'Reilly, Christina Romer, David Romer, Paul Romer, Andrew Rose, Richard Sutch, David Wise, and Janet Yellen for valuable comments. I gratefully acknowledge support from the Institute for Policy Reform, the Sloan Foundation, and the National Science Foundation for research support under grant no. SES 90-09051 administered by the Institute for Business and Economic Research at the University of California-Berkeley.

of the elderly due to inadequate savings for retirement, addiction to alcohol and drugs, criminal and gang activity, and the impact of corporate "culture" on firm performance. Economic theories of crime, savings, and organizations are deficient and yield misleading conclusions when such behavior is ignored. The behavioral pathologies that I will describe also have consequences for policies toward, for example, savings, substance abuse, and management.

Individuals whose behavior reveals the various pathologies I shall model are not maximizing their "true" utility. The principle of revealed preference cannot therefore be used to assert that the options that are chosen must be preferred to the options that are not chosen. Individuals may be made better off if their options are limited and their choices constrained. Forced pension plans may be superior to voluntary savings schemes; outright prohibitions on alcohol or drugs may be preferable to taxes on their use reflecting their nuisance costs to others; and an important function of management may be to set schedules and deadlines and not simply to establish "appropriate" price-theoretic incentive schemes to motivate employees.

I. Salience and Decisions

A central principle of modern cognitive psychology is that individuals attach too much weight to salient or vivid events and too little weight to nonsalient events. Richard Nisbett and Lee Ross (1980) describe the following thought experiment, that they consider the "touchstone" of cognitive psychology, just as the shifting of a supply or a demand curve is the central thought experiment of economics.

Let us suppose that you wish to buy a new car and have decided that on grounds of economy and longevity you want to purchase one of those stalwart, middle-class Swedish cars—either a Volvo or a Saab. As a prudent and sensible buyer, you go to *Consumer Reports*, which informs you that the consensus of their experts is that

the Volvo is mechanically superior, and the consensus of the readership is that the Volvo has the better repair record. Armed with this information, you decide to go and strike a bargain with the Volvo dealer before the week is out. In the interim, however, you go to a cocktail party where you announce your intention to an acquaintance. He reacts with disbelief and alarm; "A Volvo! You've got to be kidding. My brother-in-law had a Volvo. First, that fancy fuel injection computer thing went out. 250 bucks. Next he started having trouble with the rear end. Had to replace it. Then the transmission and the clutch. Finally sold it in three years for junk."

[quoted in Nisbett and Ross, p. 15;
from Nisbett, et al., 1976, p. 129]

The status of this additional information is only to increase the *Consumer Reports* sample by one. Mean repair records are likely to remain almost unchanged. Yet Nisbett and Ross argue that most prospective car buyers would not view the new information so complacently.

An experiment by Eugene Borgida and Nisbett (1977) confirms the intuition that salient information exerts undue influence on decisions. Freshmen at the University of Michigan with a declared psychology major were chosen as subjects. Students were asked to express preferences concerning psychology electives. Before making this decision, a control group was given only mean psychology course evaluations; others were, in addition, exposed to a panel discussion by advanced psychology majors selected so that their course evaluations corresponded to the mean. As in the Volvo thought experiment, vivid information played a greater role than pallid information; compared to the control group, those exposed to the panel chose a higher fraction of courses rated above average. To counter the argument that this bias might be due to thoughtlessness because of the unimportance of the decision, Borgida and Nisbett note that the bias was greater for those who later entered the major than for those who dropped out.

II. Procrastination

Procrastination provides the simplest example of a situation in which there are repeated errors of judgment due to unwarranted salience of some costs and benefits relative to others. In this case each error of judgment causes a small loss, but these errors cumulatively result in large losses over time and ultimately cause considerable regret on the part of the decision maker.

Let me illustrate with a personal story and then show how such behavior can be modeled. Some years back, when I was living in India for a year, a good friend of mine, Joseph Stiglitz, visited me; because of unexpected limitations on carry-on luggage at the time of his departure, he left with me a box of clothes to be sent to him in the United States. Both because of the slowness of transactions in India and my own ineptitude in such matters, I estimated that sending this parcel would take a full day's work. Each morning for over eight months I woke up and decided that the *next* morning would be the day to send the Stiglitz box. This occurred until a few months before my departure when I decided to include it in the large shipment of another friend who was returning to the United States at the same time as myself.

The preceding story can be represented mathematically in the following way. The box was left with me on day 0. At the end of the year, at date T , the box could be costlessly transported. The cost of sending the box on any day prior to T was estimated at c , the value of a day's work. I estimated Joe's valuation of the use of the contents of the box (which was the same as my value of his use of the contents) at a rate of x dollars per day. I saw no reason to attach any discount rate to his use of the box. However, each day when I awoke, the activities I would perform if I did not mail off the Stiglitz box seemed important and pressing, whereas those I would undertake several days hence remained vague and seemed less vivid. I thus overvalued the cost of sending the box on the current day relative to any future day by a factor of δ . This caused me to procrastinate.

On each day t , until date $T - c/x$, I made the dynamically inconsistent decision that I would not send the box on that day, but would instead send it the very next day. Ultimately, I decided to simply wait and send it costlessly at my departure.

Consider my decision process. On each day t , I awoke and made a plan to send the box on date t^* . I chose t^* to minimize V , the costs net of the benefits of sending the box.

If I sent the box on that day (day t), V would be

$$(1) \quad V = c(1 + \delta) - (T - t^*)x$$

for $t^* = t$.

The factor δ represents the extra salience of sending the box on that day. If I waited, but sent the box at some later time, other than the time of my departure, V would be

$$(2) \quad V = c - (T - t^*)x$$

for $t + 1 \leq t^* < T$.

And if I waited until the end of my stay to send the box, I saw that

$$(3) \quad V = 0 \quad \text{for } t^* = T.$$

On each and every day, up until day $T - c/x$, the time when the costs of sending the box just equaled the benefits of its receipt, I decided to send the box *tomorrow*. Since δc was sufficiently large, at each date t , I set the planned date for sending the box at $t^* = t + 1$. By time $T - c/x$, it was apparent that the costs of sending the box no longer exceeded the benefits, and thus I guiltily decided to ship it at the time of my return. I had procrastinated too long.

Three key features of the situation resulted in procrastination. First, the time between decisions was short. Second, in each period there was a small, but not a minuscule, "salience cost" to undertaking the job now rather than later. The condition that results in procrastination is $\delta c > x$. The daily benefit from the box, x , is small if the

time between decisions is short. δc is significant if there is a significant psychological lump sum cost to doing the project now rather than later. The third key feature of the situation was the dynamic inconsistency in my decision making. Each day I decided to put off the project until tomorrow. I did not have rational expectations, since I did not foresee that when the next day came I would continue to put off the decision for an additional day.

My procrastination was costly. The cumulative loss incurred due to my procrastinating behavior amounted to approximately $Tx - c$.¹ For each day up to the critical day at approximately $T - c/x$, I wrongly decided not to send the box. After the critical time (approximately) $T - c/x$, I made the correct decision to wait to send the box.² For every day between 0 and $(T - c/x)$, the loss from the decision made on that day was x , the cost of an extra day's use of the box. The cumulative loss was consequently the product of the daily cost of a delay, x , and the $(T - c/x)$ decisions to delay. This product is $Tx - c$, the total loss from the failure to send the box. In consequence, the cumulative cost of the errors of decision amount to the total loss that occurs over the period. Many wrong decisions all of the same type but of small value cumulated into a significant loss. And yet this loss occurred as a consequence of only a modest amount of irrationality or "undue salience."

A numerical example is useful to illustrate the necessary size of the "salience" premium δ on current relative to future work required for procrastination to occur. Suppose that I valued my time at \$50 per day and Joe Stiglitz' use of his box at 50 cents per day. If δ exceeds .01 ($= .50/50$), then procrastination will occur for 265 ($= 365 - 50/.5$) days. We consequently see that in this type of example, where there are

significant initial costs relative to the benefits, only small amounts of unwarranted salience on present relative to future action can result in significant delay.

Procrastination with Deadlines. The preceding model of procrastination has the special feature that if the task is not done in a timely fashion, it does not need to be done at all. It is like the referee's report that the editor angrily sends to another reviewer after too long a lapse. However, many tasks have deadlines. For our students, the cost of procrastination involves "pulling an all-nighter" to get the term paper (conference paper) done on time.

Qualitatively, the same type of results that we have already seen can still occur: small salience costs to beginning projects can result in costly procrastination. Consider what happens if the disutility of a project varies with the square of hours worked per day, and the number of hours to complete a project is fixed. Let the salience cost of beginning a project be a multiple of the disutility of the first day's work. In an example in which the salience cost is 2 percent of the total cost and the length of the project is 100 days, the added cost of completing the project can be calculated at approximately 41 percent.³

³Let us suppose that the daily utility cost of doing a project varies with the square of the number of hours worked per day, and that the project, without procrastination, would require Th hours of labor. Then we can write the intertemporal utility function as $U = \sum_{t=0}^T e_t^2$, where e_t is the number of hours worked on day t . Without procrastination, the total utility cost of the project is $U = Th^2$.

Let us now compare this to the cost borne by a procrastinator. For the procrastinator, current costs are unduly salient in comparison with future costs. The salience premium is δh^2 , a multiple of δ of the daily cost of the project if begun on time. The perceived cost of completing the project, if begun at date τ , is thus:

$$V = \delta h^2 + \sum_{t=\tau}^T e_t^2.$$

In each period, the procrastinator compares the total cost V of the project if begun that day (including the added salience cost δh^2 of that day's input) with the cost of waiting one more day to begin, taking no

¹The exact loss is $Tx - c(1 + \delta)$. If I sent the box on date 0, V has value $c(1 + \delta) - Tx$. Since I sent the box at T , $V = 0$. The difference is $Tx - c(1 + \delta)$.

²The exact critical date is the first day on which the decision maker decides to send at T : i.e., the smallest t such that $c - (T - t - 1)x > 0$.

It may also be worth noting that if the salience value of beginning the project increases with the intensity of the first period's work, a project, such as a reducing diet, may never be begun, or a task may be begun at the latest possible date at which completion is feasible.

III. Procrastination: Substance Abuse, Savings, and Organizational Failures

At first glance, my examples of procrastination may appear to be of no relevance to economics. However, I want to argue that such behavior may be critical in understanding the causes of such varied problems as drug abuse, inadequate savings and some types of organizational failure.

A. Substance Abuse

It has often been observed that consumers are knowledgeable about their decisions, and that their decisions are utility maximizing. Ethnographies of drug users suggest that drug use is no exception. Gary Becker and Kevin Murphy (1988) and George Stigler and Becker (1977) have developed the theory of such behavior in their forward-looking models of rational addiction. In these models, use of a good affects the future enjoyment of its consumption, but people correctly foresee these changes in taste. The application of such models, combined with utilitarian ethics, leads to the conclusion that drug use should be legalized with a tax reflecting its nuisance to others.

account that when the next day arrives, it too will have special salience. Behaving in this way, the project is begun with approximately $T\delta^{-1/2}$ days left for its completion. This is a poor approximation for low values of δ because for low values of δ , it does not pay to procrastinate. This increases the total cost of the project by a multiple that is the square root of δ . In this example, a small salience cost of beginning a project results in losses from future actions that are a multiple of the cost. For example, if T is 100 days and δ is 2, the salience cost of beginning the project relative to the total nonprocrastinating cost is only .02. But the total cost of completing the project increases by 41 percent ($\sqrt{2} - 1$).

I do not agree with this conclusion, because I do not agree that the model of forward-looking, rational behavior accurately describes the way in which individuals decide on drug or alcohol intake. Most drug abusers, like most chronically overweight individuals, fully intend to cut down their intake, since they recognize that the long-run cost of their addiction exceeds its benefits. They intend to stop—tomorrow. Individuals following the procrastination model are both maximizing and knowledgeable, and yet their decisions are not fully rational. For example, psychologist Roger Brown describes addictions in the following way.

Actions like smoking a cigarette, having a drink, eating a candy bar, and working overtime to "catch up" all lead to immediate and certain gratification, whereas their bad consequences are remote in time, only probabilistic, and still avoidable now. It is no contest: Certain and immediate rewards win out over probabilistic and remote costs, even though the rewards are slight and the possible costs lethal. [1986, p. 636]

Ethnographies of drug abusers reveal that most are well aware of the long-term consequences of their habit and many have detailed and subtle knowledge of the effects of drugs. (See, for example, Cheryl Carpenter et al., 1988 and Harvey Feldman et al., 1979.) They apply this knowledge to avoid some of the worst potential consequences of drug use. An interview with Don, an "angel dust" (PCP) user in the Seattle-Tacoma area reveals the knowledge of the long-term effects of drug use, and also an inability to use the knowledge to quit. Don tells his interviewer:

And every time I use the drug and I come down and I am straight, since I do have this heightened form of awareness and perspective, I always tell myself, "Well, that's the last time I'll use it. I don't need it now." I can see where this is, what I've got to do, and this is what I want to do and everything falls into place.

[Feldman et al., p. 137]

Later, I will discuss some ways in which the social pressures emanating from group dynamics reinforce individual reasons for addiction.

B. Savings

The procrastination model may also pertain to intertemporal savings and consumption decisions. The modern textbook/journal article model of consumption and savings decisions typically views agents as maximizing a time-separable utility function with discount rate δ . This discount is said to parameterize agents' impatience. Curiously, economists who build models with utility functions of this sort consider themselves to be modeling the behavior of *rational* consumers. Yet early discussion of impatience viewed discounting as *irrational* behavior. Irving Fisher regarded such impatience as evidence of lack of foresight or lack of will. In this regard, he writes,

Generally speaking, the greater the foresight, the less the impatience and *vice versa* This is illustrated by the story of the farmer who would never mind his leaking roof. When it rained he could not stop the leak, and when it did not rain there was no leak to be stopped! Among such persons, the preference for present gratification is powerful because their anticipation of the future is *weak*. [1930, p. 81]

Fisher's example of the farmer fits the model of an agent continually making an inconsistent decision.

A clear moral of the procrastination model is that time-inconsistent behavior is especially apt to occur when there is some fixed cost (perhaps not very great) to beginning a task, the "periods" are short, and the per period cost of delay is low. Many personal financial decisions satisfy these conditions. A good example concerns the behavior of junior faculty at Harvard. Due to some institutional oddity, university contributions to TIAA/CREF cumulated without payment of interest until the recipient filed a form indicating his desired allocation between the two retirement funds. This could

be done at any time and took less than an hour. And yet most junior faculty who left Harvard in the 1970's made this decision only upon their departure. They thus lost hundreds of dollars worth of interest by failing to do an hour's work.⁴

A more serious application of the procrastination model is to savings.⁵ Most of the U.S. elderly (those over 65) derive relatively little of their income from financial assets. In 1971, 51 percent of the elderly had no income from financial assets; 78 percent had less than 20 percent of their income from financial assets (Michael Hurd, 1990, p. 571). This stark absence of financial asset income is consistent with the hypothesis that most households would save very little, except for the purchase of their home and the associated amortization of mortgage principal, in the absence of private pension plans. Partly because these additions to financial assets are so small, some seemingly paradoxical results have been obtained. Phillip Cagan (1965) and George Katona (1965), for example, find a *positive* relation between pension plans and private saving. In the life cycle model (up to some limit), \$1 of pension savings should lead to \$1 reduction of private saving. Steven Venti and David Wise (1986, 1987) report results similar to those of Cagan and Katona. They find no significant relation between ownership of a pension plan and willingness to invest in IRAs. Alicia Munnell's (1976) findings are less extreme. She looked at the savings of a sample of 5,000 men aged 45 to 59 in 1966. She estimated that a \$1 private

⁴I owe this observation to Janet Yellen.

⁵Richard Thaler and Hershey Shefrin (1981) discuss the role of Christmas Clubs in forcing the scheduling of savings. Their model of saving behavior, and of procrastination, is different from my model in this lecture. Their model discusses two types of decision making: for long-term planning and for maximization of current utility. People may constrain themselves (i.e., may make arrangements such as Christmas Clubs) so that they can then be free to maximize their short-term utility without further constraint. In this way budgets act as mental accounts. The Christmas Clubs relative to my model set clear schedules for saving, which result in penalties if not followed, and thereby prevent procrastination.

pension contribution caused a reduction in nonpension savings of 62¢ for these men nearing retirement. This is still considerably less than the \$1 that would be expected if people were life cycle savers and pension plans did not induce oversaving.⁶

The hypothesis that, in the absence of pension plans, many individuals lack sufficient self-discipline to begin saving for retirement in a timely fashion is consistent with the finding that there were high rates of elderly poverty prior to the rapid, unexpected growth in Social Security payments in the late 1960's and the 1970's. In 1966, the elderly poverty rate was 30 percent, fully double the poverty rate of the nonelderly (David Ellwood and Lawrence Summers, 1986, p. 81).⁷

C. Organizational Failures

Procrastination is as prevalent in the workplace as in the home. Procrastination by workers results both in delay in initiating projects that should be begun as well as in delay in terminating projects that should be ended.⁸

⁶It is also low because we might expect those without pension plans to be making up for prior failure to save as they near retirement, just as the procrastinating student has to work especially hard near his term paper deadline.

⁷This high rate may reflect the prior lives of poverty of the elderly population in 1966; this group spent much of their working lives in the Great Depression. But earlier statistics, from such indicators as the fraction of elderly living in poorhouses, show that the elderly had particularly high poverty rates in the 1920's, before both modern pension plans and the Great Depression (Michael Dahlin, 1983).

⁸In their advice book on *Procrastination*, Jane Burka and Lenora Yuen (1983) urge potential procrastinators to set clear and realistic schedules for themselves and then adhere to them. In the preceding Stiglitz-box model, the determination of a schedule that was binding would have resulted in the box being sent on day 1 or day 2. Thomas Schelling (1985, p. 368) has explained in similar terms why parents at the beach may give their children the clear advice not to go in the water at all, even though they do not mind the children getting a little bit wet. In the absence of a clear "schedule" telling the children when the water is too deep, they may wade ever deeper and end up in danger.

In private life, individuals are frequently forced to self-monitor their behavior, as in stopping an addiction, writing a Ph.D. thesis, devising a private asset plan, or sending off referee reports; in such areas procrastination can easily occur leading to serious losses. However, in work situations, outside monitoring is possible, and a major function of management is to set schedules and monitor accomplishment so as to prevent procrastination.

Proper management not only prevents procrastination in project *initiation*; it also prevents procrastination in project *termination*. Psychologists have found tendencies to delay terminations of projects by people who consider themselves responsible for their initiation. Barry Staw (1976) divided a group of 240 undergraduate business students into two groups. One group was asked to decide on investment allocation in a business school case study of the Adams and Smith Corporation. They were then asked to make a further allocation of investment between the division with their initial project and the other division of the firm. In contrast, a control group only had to make the second allocation. Both groups, however, had matched past histories of the firm and the success of the firm's projects. In the case of project failure, those who had made a prior commitment wanted to invest significantly more in that division than the control group with the same matched history who had made no such prior commitment. One explanation matches our model: that failure to terminate the project puts the painful decision off until tomorrow; the pain of admitting a mistake today is salient relative to the pain of having to admit a possibly even bigger mistake tomorrow. This same phenomenon may also be explained, not necessarily inconsistently, by cognitive dissonance. Once people have made decisions, they avoid information that does not support that decision because it is psychologically painful.

Staw and S. McClane (1984) report how the commercial division of a large American bank avoids procrastination in loan cutoff decisions. Loan officers are not penalized for cutting off loans, although they are pe-

nalized for failing to foresee possible losses. They are especially penalized if loan losses are discovered by bank examiners before they are reported. Most important, loans with significant difficulties are referred to a separate committee not involved in the initial decision to obtain the maximum salvage value.

In the next section I will discuss how courses of action are reinforced by selective elimination of information contrary to that course of action, so that initial psychological overcommitments are reinforced. Jerry Ross and Staw (1986) examine the history of Expo 86 in Vancouver, whose projected losses escalated from a maximum of \$6 million to over \$300 million. In this case, exit costs reinforced the initial psychological overcommitment: the Prime Minister of British Columbia feared loss of election on termination, contracts would have to be breached, and outside vendors would make losses from investments made in anticipation of the fair.

Staw and Ross (1987) list management practices that limit overcommitment: administrative turnover; low cost to executives for admitting failure; lack of ambiguity in data regarding performance; allowing losses to be blamed on exonerating circumstances; separating termination decisions from initiation decisions; and considering from the beginning the costs and procedures of project termination.

IV. Indoctrination and Obedience

Irrational obedience to authority is a second type of "pathological," time-inconsistent behavior with important social and economic ramifications. Procrastination occurs when there is a fixed cost of action today and current costs are more salient than future costs. Undue obedience to authority may occur as a form of procrastination if disobedience of an authority is salient and distasteful. In addition, authority may be particularly powerful when yesterday's actions affect the norms of today's behavior. Both such influences (the salience of current disobedience and a shift in the utility of subjects in accordance with their prior actions) are present in Milgram's experiments, which I shall review.

The subjects in the Milgram experiment were adult males, recruited by a mail circular requesting participation in an experiment purportedly concerning the effects of punishment on memory. The subjects were assigned the role of teacher, while an accomplice of the experimenter, a professional actor, played the role of learner. The subjects were instructed to administer shocks to the learner when he gave wrong answers. The shocks were a learner-discipline device. The learner, a trained actor instructed to simulate the appropriate reactions to the shocks administered by the subjects, was visible to the subject through a glass window, and, unbeknownst to the subject, was not wired. Subjects initially gave low voltage shocks (15 volts) with doses increasing 15 volts at a time to a maximum of 450. There are different versions of this experiment, but, in all versions, the learner showed significant response to the shocks. For example, in one version, according to Milgram's description, "at 75 volts the learner began to grunt and groan. At 150 volts he demands to be let out of the experiment. At 180 volts he cries out that he can no longer stand the pain. At 300 volts he ... [insists] he must be freed" (1965, p. 246, quoted in E. Stotland and L. K. Canon, 1972, p. 6). Despite these protests by the learner, 62.5 percent of the subjects reached the maximum of 450 volts. The experiment has been repeated under a wide variety of conditions, but always with the same result: a significant fraction of the population administers the maximum dosage.

As important as the primary finding from Milgram's experiment that individuals are remarkably obedient is the further finding of their lack of awareness of this trait in themselves and in others. Elliot Aronson (1984, p. 40), a professor of social psychology at UC-Santa Cruz, asks the students in his classes how many would continue to administer shocks after the learner pounds on the wall. Virtually no one responds affirmatively. Milgram conducted a survey of psychiatrists at a major medical school who predicted that most subjects would not administer shocks in excess of 150 volts, and virtually no subjects would administer the maximum 450 volts. This finding supports

my central argument: that in appropriate circumstances, people behave in time-inconsistent ways that they themselves cannot foresee, as when they procrastinate or exhibit irrational obedience to authority.

A Model of Behavior in the Milgram Experiment. Let me present a simple model that is a variant of the previous model of procrastination and which explains the sequential decisions made by the subjects in Milgram's experiment. I shall first assume that current disobedience by the subject is especially painful; it is especially salient. Lee Ross (1988) has argued that special salience is attached to disobedience because there is an implicit contract between the teacher and the experimenter. The experiment defines the situation so that there is no legitimate way for the teacher to terminate.⁹ Thus the subject sees the cost of current disobedience as very high, although in an ill-defined way, he may plan to disobey in the future. Second, I shall assume that the subject suffers a loss in utility, not based on the current voltage he administers to the learner, but instead on the deviation of the current voltage from what he last administered. (Alternatively his utility might depend on the deviation from the highest voltage previously administered.) This model is consistent with cognitive dissonance. Once people have undertaken an action, especially for reasons they do not fully understand, they find reasons why that action was in fact justified. In this formulation, the subject decides to obey up until time T so as to maximize V_t .

If he disobeys today at time t , his utility is

$$(4) \quad V_t = -bD(1 + \delta).$$

But if he postpones obeying, his expected utility is

$$(5) \quad V_t = -bD - c \sum_{k=t}^{T-1} (W_k - W_{t-1})$$

$T \geq t + 1$

⁹Ross suggests that if teachers had a red button to push that would allow them to stop the experiment, very few subjects would have given the maximum dosage. In my model, this would decrease the value of δ , the special salience attached to current disobedience.

if he first disobeys at time $T \geq t + 1$, where δ is the extra salience attached to today's disobedience, D is the cost of disobedience, W_k is the voltage of the shock administered at time k , and W_{t-1} is the norm for the level of shocks determined by previous actions.

It can easily be seen in this formulation that at each date, with sufficiently slow expected future escalation of commands the subjects can be led, as in the Milgram experiment, to deliver ever-higher levels of shock. They plan to disobey in the future if the experiment continues, but not currently. While planning future disobedience if escalation continues, at the same time these subjects are continuing to raise the level of shock required to induce them to disobey. The dependence of norms of behavior on previous actions does not just cause continued poor decision making due to postponement, but also causes escalating errors in decisions.

While V may be the function that subjects maximize in the heat of the moment, under the conditions of the Milgram experiment, a more accurate expression of their *true* intertemporal utility function might be

$$(6) \quad V_0 = \sum_k \{ -bD_k - cW_k \}$$

where V_0 is their intertemporal utility and k sums over all the trials.

Such a utility function is reflected in the postexperiment interviews and follow-up questionnaires. Most of the subjects were, in retrospect, extremely regretful of their decisions in the experiment. For example, one subject, who was a social worker, wrote in a follow-up questionnaire a year later:

What appalled me was that I could possess this capacity for obedience and compliance to a central idea, i.e. the value of a memory experiment, even after it became clear that continued adherence to this value was at the expense of the violation of another value, i.e. don't hurt someone else who is helpless and not hurting you. As my wife said, "You can call yourself Eichmann."
[Milgram, 1975, p. 54]

The preceding models of procrastination and obedience concern actions that occur because individuals possess cognitive structures of which they are less than fully aware. The assumption that such structures influence behavior is unfamiliar in economics, but central to other social sciences. A major task of psychology is to discover such unperceived behavioral regularities; the concepts of *culture* in anthropology and *the definition of the situation* in sociology both concern cognitive structures only dimly perceived by decision makers.

The Milgram experiment demonstrates that isolated individuals can exhibit remarkably obedient (and deviant) behavior inside the laboratory. In group situations, however, there is evidence that such behavior occurs only when there is near unanimity of opinion. In this regard, the most relevant evidence comes from a variant of the Asch experiment. Solomon Asch (1951, p. 479) found that subjects asked to match the length of a line to a comparison group of lines of different length gave the wrong answer roughly 40 percent of the time if they were preceded by confederates of the experimenter who had previously given the wrong answer. However, in another variant of the experiment, Asch (1952) found that the presence of just a single confederate who gave the right answer in a large group of confederates reduced the number of wrong answers by a factor of two-thirds. This suggests that the presence of like-minded others significantly raises the likelihood of disobedience in situations such as the Milgram experiment. It might be inferred that obedience such as obtained by Milgram could only occur in the laboratory where people are shielded from outside information and influences.

The next four sections will present examples of individuals who participate in groups and make regrettable decisions. In each of the examples, a sequence of small errors has serious ill consequences. Furthermore, in each of the situations described, there is a natural equilibrium in which those who disagree with the actions taken find it disadvantageous to voice their dissent, which is accordingly isolated from the decision-making process.

V. Cults

A. Unification Church

Evidence seems to show that neither members nor inductees into cult groups such as the Unification Church (Moonies) are psychologically very much different from the rest of the population (see Marc Galanter et al., 1979; and Galanter, 1980). The method of induction into the Moonies indicates how normal people, especially at troubled times in their lives, can be recruited into cults and the cult can persist. Membership into the Moonies involves four separate decisions. Potential recruits are first contacted individually and invited to come to a 2-day, weekend workshop. These workshops are then followed by a 7-day workshop, a 12-day workshop, and membership. The potential recruit in consequence makes four separate decisions: initially to attend the 2-day workshop, to continue into the 7-day workshop, and then again into the 12-day workshop, and finally to join the Church. As in the Milgram experiment, the membership decision is achieved in slow stages.

Consider the process from the point of view of the potential recruit. Those who agree to attend the first 2-day workshop must have some predisposition toward the goals of the Church; otherwise they would not have attended. But they are probably surprised on arrival to find so many like-minded persons. In addition, the members of the Church intermingle and apply gentle persuasion in the first 2-day workshop; the inductees' commitment at this point begins to change. Then, continuing with the 7-day workshop, and again with the 12-day workshop, only the most committed continue; those who disagree leave. At each stage the Church members are thus able to increase the intensity of their message. As in the Milgram experiment and other social psychology experiments on conformist behavior, the potential inductee, in the absence of disagreement, is likely to change his opinions. And, as we have seen, because of the self-selection process, there is unlikely to be strong disagreement among the workshop attendees. Galanter's study of eight workshop sequences reveals this gradual attrition

according to commitment. Of the 104 guests at the initial 2-day workshops, 74 did not continue. Of the 30 at the 7-day workshops, 12 did not continue (including a few who were rescued by their families, and a few who were told not to continue by the Church). Of the 18 remaining at the 21-day workshops, 9 did not continue to membership. And of the remaining 9, 6 were active church members 6 months later.

The example of the Moonies illustrates a process of conversion. Converts make a sequence of small decisions to accept authority. Ultimately, as a result of this sequence of decisions to obey rather than to rebel, the converts develop beliefs and values very different from what they had at the beginning of the process. This willingness to acquiesce to authority is abetted by self-selection. Those who agree most with the Church self-select into the group. Because those who disagree most exit, the dissent necessary for resistance to escalation of commitment does not develop.

B. Synanon

The case of Synanon is possibly the best studied, as well as being, in the end, one of the most horrific of these cult groups. In this group we can see the pressure for obedience to authority operative in the Milgram experiment, as well as the selective exit of would-be dissenters who could break the isolation necessary to maintain this obedience to authority.

Synanon was initially an organization devoted to the cure of drug addicts, but it gradually evolved into a paramilitary organization carrying out the increasingly maniacal whims of its founder and leader. (My account comes from David Gerstel, 1982.) The leader, Charles Dederich, as well as the other founders of Synanon, adapted the methods of Alcoholics Anonymous for the treatment of drug addiction. At the time, little was known about drug abuse in this country; it was also widely believed that drug addiction was incurable. By proving the contrary, Synanon received considerable favorable publicity. With aggressive solicitation of gifts (especially of in-kind tax deductible gifts) and commercial endeavors

such as the sale of pens, pencils, and briefcases with the Synanon logo, it expanded from a houseful of ex-addicts, first to a large residential site in Santa Monica, and, at its peak, to several residential communities in both northern and southern California with more than 1,600 residents (Richard Ofshe, 1980, p. 112).

To understand the path of Synanon from these benign origins into what it eventually became, it is necessary to focus on the methods of control in the organization. The members led dual lives: daytime workday lives, and nighttime lives spent in a pastime called *The Game*. The daytime lives of members were devoted to hard work, especially for the cause of the community. Members were given virtually no private property and were expected to donate their own resources to Synanon; they had virtually no privacy. Gerstel reports his first impressions of Synanon as amazement at the cleanliness of the buildings, the orderliness of the grounds, and the cheerfulness of the workers. The daytime code of Synanon was to maintain a cheerful positive attitude at all times, exemplified by the song of the trashmen: "We're your Synanon garbage men./ We don't work for money. Ooooh, we don't work for cash./ We work for the pleasure/ Of taking out yo' trash" (Gerstel, p. 5).

At night, however, the unbridled positivism was given up, and members acted out their hostility and aggressions in *The Game* (adapted from the practices of Alcoholics Anonymous, from which Synanon originated). Participants in the game were expected to be brutally frank in criticizing any other member who did not live up to the standards expected of Synanon. Because the lives of the members were so open to each other, these criticisms could extend to the smallest detail. Since members had virtually no privacy, this criticism naturally monitored any deviation from the purposes of the organization. The incentives of *The Game* induced members to strive to maintain the very best behavior.

The Game, however, like any other game, had rules, and those rules led to complete control of this fairly large community by its leader. These rules encouraged the criticism of members by one another, but forbade

criticism of the goals of Synanon itself or any shortcomings of its leader. Anyone who criticized the organization or its leadership would be harshly criticized by all (an incentive not to engage in such activity). If that criticism persisted, the offender would be banished from the community.

Under these rules of behavior, Synanon evolved into an organization under the control of a leader who became increasingly insane. In the late 1970's, Dederich insisted that members follow his every whim that included, to give some examples, enforced dieting (the Fatathon), enforced vasectomies for all males, and an enforced change of partners, first of all for married members of the community, and subsequently for all paired members whether married or not. Those who did not go along with these measures were "gamed," that is, criticized vehemently in *The Game*, beaten up, or evicted. During this period, Dederich was also building up his own armed paramilitary force that reacted against threats both within and outside the community. Within Synanon, dissenters were beaten up. Outside, passersby or neighbors with real or presumed insults aimed at the community were accosted and beaten, often severely. One former member, who was suing for the custody of his child still living there, was beaten to the point of paralysis, never to recover. Dederich was eventually convicted on a charge of conspiracy for murder—for sending unwanted mail to a Los Angeles attorney who was fighting Synanon: two Synanon vigilantes were found leaving a poisonous rattlesnake in his mail box.

The Synanon experience follows closely what Milgram observed in the laboratory. At each move by Dederich, the members were forced individually to decide whether to obey or to disobey. Disobedience involved the present cost of leaving the group and seeking immediately a new way of life with insufficient human and financial resources. Many members in the past had found life outside Synanon painful and had sought refuge there. Thus the consequences of disobedience were immediate and salient. As members chose the course of obedience, their norms of behavior as members of Synanon gradually changed, just as the

norms of behavior of Milgram's subjects changed according to the level of punishment they had previously administered. The process was aided, in Synanon as in Milgram's laboratory, by the absence of dissent. In Synanon the absence of dissent was ensured in usual circumstances by *The Game* and in unusual circumstances by forced expulsions.

VI. Crime and Drugs

Economists modeling crime (see Becker, 1968) and drug addiction have viewed the decisions to engage in these activities as individually motivated. Becker and Murphy, following Stigler-Becker, have even viewed the decision to pursue addictive activities as both rational and forward looking. The Milgram experiment and the behavior of cult groups, if nothing else, serve as warnings. It is inconceivable that the participants in Milgram's experiment were forward looking. These participants could not imagine that anyone (least of all themselves) would behave as they ultimately did. Likewise, the flower children of Synanon of the 1960's could not have conceived of themselves turning into gun-toting toughs in the 1970's. The assumption of forward-looking rationality regarding the change in their consumption capital, to use Stigler-Becker terminology, is totally violated.

The analogy between cult groups and the behavior of teenage gangs, where most criminal activity and drug addiction begin, is fairly complete. A member of a teenage gang typically finds himself (much less frequently herself) in a position very similar to that of a member of Synanon. The typical member of a gang makes a sequence of decisions that results in an escalating obedience to the gang leadership. At each stage of his career as a gang member, he makes the choice whether to obey or to disobey. In extreme cases, disobedience leads to expulsion from the gang. The gang member thus faces the same dilemma as the member of Synanon: whether to forsake friends who are close and in an important respect define his way of life, or to go along with the gang decision. In rising from junior to senior membership in the gang, or in following a

leader who himself is becoming deviant, the gang member by obeying increases his commitment to obedience. The situation is exactly analogous to that of subjects in the Milgram experiment.

Furthermore, the isolation from dissent obtained in Milgram's laboratory also naturally occurs in teenage gangs. The major activity of such gangs, according to David Matza (1964) is hanging out, and the major activity while hanging out is insulting other gang members to see how they respond. This activity is called "sounding," because it measures the depths of the other member on his response to the insult. The depth probe usually focuses on the manliness of the gang member and/or his commitment to the gang itself. The probing of his commitment to the gang plays the same control function as *The Game* in Synanon. Those who display less than full commitment to the gang in sounding, or to Synanon in *The Game*, suffer a form of public censure. Such procedures make members reluctant to voice their disagreements with the goals or activities of the gang, just as members of Synanon found it difficult to display negative attitudes toward the group. Thus members of teenage gangs find themselves in isolated positions, unable to resist the aims of powerful and deviant leaders. The ethnographies we have of such gangs support the importance of sounding and the role of important leaders who play a disproportionate role in planning gang activities (see Jay MacLeod, 1988, and William Whyte, 1943).

Just as the participants in the Milgram experiment "drifted" into obedience, and members of Synanon drifted into gangsterism, Matza (1964) showed how teenagers "drift" into delinquency. Matza (1969) likens the process of becoming delinquent to "religious conversion." The analogies of *drift* and *conversion* are both consistent with my model of time-inconsistent behavior. Like the cult groups I have just described, delinquent teenage gangs have mechanisms that work to preserve their isolation from outside influences. Should we be surprised that many such gangs with few social constraints engage in harmful deviant activity?

Consider the activities, as chronicled by MacLeod, of the "Hallway Hangers," a gang

who live in a low-income housing project in a New England city. The major activity of this gang is hanging out in Hallway #13 and sounding each other, with varying degrees of playfulness and malice. While hanging out, the gang ingests a wide variety of stimulants, including beer in vast amounts, a great deal of marijuana, some cocaine, PCP, and mescaline, and occasionally some heroin. The central value of this group is its loyalty to the gang and the other members, just as Synanon's central value from its inception to its end was the loyalty of members to the general community. The ethos of this gang is illustrated by MacLeod's story of Shorty and Slick, when they were caught ripping off the local sneakers factory (as told by Shorty):

See, that's how Slick was that day we were ripping off the sneakers. He figured that if he left me that would be rude, y'know. If he just let me get busted by myself and he knew I had a lot of [...] on my head, that's what I call a brother. He could've. I could've pushed him right through that fence, and he coulda been *gone*. But no, he waited for me, and we both got arrested. I was stuck. My belly couldn't get through the [...] hole in the fence.

[pp. 32-33]

This same aspect of gang behavior was emphasized 50 years ago in the classic street corner ethnography by Whyte. He explained the lack of social mobility of the most capable corner boys by their unwillingness to adopt a life style that would have sacrificed friendships with peers who would not advance with them. Just as Slick did not run when Shorty got caught in the fence in MacLeod's account, the leader of the "corner boys" Doc, in Whyte's account, fails to advance himself in school so he can remain with his friends.

Such gangs provide a perfect social environment for regrettable decisions. Gang members find the costs of nonacquiescence especially salient, since such nonacquiescence leads to isolation from the social group to which they are committed. As occurred at Synanon in a similar environment, gang members can then be led step-by-step

to escalating levels of crime, drugs, and violence, with each preceding step setting the norm for the next.

The question remains how to alter such behavior by social policy. William Wilson (1987) has argued the importance of the move from the central city to the suburbs of the middle class, which, he says, has resulted in the disappearance of social networks that formerly were the pathways to employment. According to Wilson, the result, especially in the black community, has been a dearth of employed (and therefore eligible) males and a dramatic increase in out-of-wedlock births.

There is, however, another effect of the disappearance of the urban middle class for poor youth left in the central cities. This disappearance has left fewer alternative social groups for those who do not want to acquiesce in the violent acts of their peers, thus making such acquiescence and gang violence more frequent.

Social policy should have the role of recreating, artificially if necessary, the beneficial social networks that have vanished. This would reduce the cost of dissent by gang members to criminal, violent, or drug-prone actions by providing alternatives. In addition, we have seen that just a little bit of dissent, and therefore perhaps just a little bit of information, may stop escalation toward commitment. Lisbeth Schorr (1989) has compiled a long list of social projects that have significantly reduced the problems of the underclass, problems such as teenage pregnancy, school truancy, drug abuse, violence, and alcoholism. In each of these projects, the key to success has been the special effort by social workers involved to gain the trust of their clients. The success of these projects shows that, when isolation can be broken and trust established, small amounts of information can significantly reduce the number of regrettable decisions.¹⁰

Evidence for the view that social isolation results in high crime rates comes from the

positive correlation between crime rates and city size. Smaller cities have less room for specialization in social groups than larger cities—so that isolation from common social norms is more difficult to attain. They also have lower crime rates. Cities with less than 10,000 people have one-fifth the violent crime rates of cities with populations more than 250,000. In 1985, cities with over 250,000 people had 50 percent higher violent crime rates than those cities with populations between 100,000 and 250,000 (U.S. Bureau of the Census, 1987, p. 157).¹¹

VII. Politics and Economics

Economists who have applied the tools of their trade to the political process have studied the workings of democracy and majority rule under individualistic values (see Kenneth Arrow, 1963). They are optimists. The model of cult group behavior, in contrast, is relevant in understanding politics' darker side. I will give two illustrations.

A. *Stalin's Takeover*

My first example concerns Stalin's ascension to power in Russia. The history of the Bolshevik party and the history of Synanon are strikingly similar. (I take the Bolshevik history from Isaac Deutscher, 1949). Initially, there were the early days of reformist zeal, of meeting secretly in lofts, warehouses, and other strange places. But, in addition, and most importantly, there was commitment to the organization. To the Bolsheviks, this commitment was of paramount importance. Indeed, it was over the constitutional issue as to whether party members should merely be contributors (either financial or political) or should, in addition, submit to party discipline that split the Russian socialist workers' movement into two parts—the Mensheviks and the Bolsheviks. This loyalty to party discipline, useful in the revolution, ultimately enabled

¹⁰In the case of the Unification Church, such active intervention frequently occurred as members were captured by relatives and forcibly deprogrammed (Galanter, 1989).

¹¹Some of these differences undoubtedly are due to the concentrations of poor people with high crime rates in central cities that are large in size.

Stalin to take over the party and pervert its ideals. It underlay the acquiescence of his tough comrade revolutionaries in the scrapping of the original principles of Bolshevism: open intraparty debate and dedication to the cause of the workers and peasants. In the 1920's and 1930's, as Stalin collectivized the peasants and tyrannized over dissidents, these old comrades stood by, perhaps not quite agreeing, but not actively disagreeing either, much like Milgram's passively obedient, passively resistant subjects. Even Trotsky in exile did not unambiguously oppose Stalin until the purges had begun as a series of decisions were made that increasingly brutalized the peasantry and cut off political debate. The exception to the lack of dissent proves the rule. Nadia Alliluyeva was the daughter of one of the founding Bolsheviks and thus an heiress by birth to the ideals of the party. She was also Stalin's wife. When Stalin collectivized the peasantry, moving perhaps 80 million from their farms in six months' time, she voiced her disapproval at a party—he replied savagely. That night she committed suicide.¹² This behavior contrasts with the party leadership who, like Milgram's subjects, had been participating in the decisions that were being taken. At each juncture, they were confronted by the decision whether to break ranks with the increasing brutalization of the peasants and the choking off of dissent, or to remain loyal to the party. By acquiescing step by step to the crescendo of Stalin's actions, they were committing themselves to altered standards of behavior. In contrast, Nadia Alliluyeva, who had withdrawn from the decision-making process to be wife and mother, could feel proper revulsion at the deviation of the party's actions from its prior ideals.

B. Vietnam War

A second example of the type of deviant group process I have described occurred in President Johnson's Tuesday lunch group,

which was the executive body controlling U.S. military decisions in the Vietnam War (Irving Janis, 1972). Here we see all of the features characterizing our model of salience, authority, and obedience. First, there was the gradual escalation of violence against the Vietnamese. Bill Moyers, reflecting on Vietnam policy after he was out of office, precisely describes how this escalation of commitment happened: "With but rare exceptions we always seemed to be calculating the short-term consequences of each alternative at every step of the [policymaking] process, but not the long-range consequences. And with each succeeding short-range consequence we became more deeply a prisoner of the process" (Janis, p. 103). The subjects in Milgram's experiments could have said exactly the same thing.

The control of dissension within President Johnson's Tuesday lunch group bore close resemblance to the processes at work in Synanon and the Hallway Hangers. The president would greet Moyers as "Mr. Stop-the-Bombing"; similar epithets were applied to other dissenters within the group: "our favorite dove," "the inhouse devil's advocate on Vietnam" (Janis, p. 120). A teenage gang would probably consider these soundings mediocre, but their lack of style may not have affected their impact. And the measures within the group which were taken to enforce unanimity ("groupthink" according to Janis) were supplemented by more or less voluntary exit as dissenters at different times came to disagree with the policy: Bill Moyers, George Ball, McGeorge Bundy, and Robert McNamara. Interestingly, since each of these individuals exited fairly soon after they developed deep reservations about the policy, there was active dissent for only a small fraction of the history of the group.

VIII. Bureaucracies

My examples of obedience to authority, so far, have centered primarily on noneconomic phenomena: religion, crime, drugs, and politics. However, the phenomenon of obedience to authority is also prevalent in bureaucracies that in a modern industrial-

¹²An alternative account of Alliluyeva's decision to commit suicide and Stalin's activities on the night before is given by Nikita Khrushchev (1990). The two accounts are not mutually exclusive.

ized society, are the sites of most economic activity.

One function of bureaucracies, following Robert Merton (1968, p. 250) and Weber, is to create specialists. A second function of bureaucracies is "infusing group participants with appropriate attitudes and sentiments" (Merton, 1968, p. 253). We could interpret the Milgram experiment as a toy bureaucracy, and my model of that experiment as a model of that bureaucracy. In that case, W_{t-1} , the level of voltage that a subject has grown accustomed to giving, constitutes his "attitudes and sentiments." In Merton's terms, it defines his bureaucratic personality.

The specialization mentioned earlier can result in bureaucratic personalities that are "dysfunctional," to use Merton's terminology. We have already seen such dysfunction in the behavior of the Moonies, Synanon, teenage gangs, drug and alcohol abusers, the Bolshevik party, and President Johnson's Tuesday lunch group. The changes that occurred in individual decision-making behavior were "latent," to use another of Merton's terms, since they were not understood by the participants and were unintentional. Furthermore, these changes occur exactly as I have been picturing: in making a sequence of small decisions, the decision maker's criteria for decisions gradually change, with preceding decisions being the precedent for further decisions. Merton gives an example of the consequences of such bureaucratically engendered personalities in the U.S. Bureau of Naturalization concerning the treatment of the request for citizenship of Admiral Byrd's pilot over the South Pole.

According to a ruling of the Department of Labor Bernt Balchen... cannot receive his citizenship papers. Balchen, a native of Norway, declared his intention in 1927. It is held that he has failed to meet the condition of five years' continuous residence in the United States. The Byrd antarctic voyage took him out of the country, although he was on a ship carrying the American flag, was an invaluable member of the American expedition, and in a region to which there is an American claim because of the explo-

ration and occupation of it by Americans, this region being called Little America.

The Bureau of Naturalization explains that it cannot proceed on the assumption that Little America is American soil. That would be trespass on international questions where it has no sanction. So far as the bureau is concerned, Balchen was out of the country and technically has not complied with the law of naturalization. [p. 254, quoted from *The Chicago Tribune*, June 24, 1931, p. 10]

Popular proponents of bureaucratic reform (for example, William Ouchi, 1981, and Thomas Peters and Robert Waterman, 1982) have emphasized the benefits of non-specialization within firms precisely because they recognize that nonspecialists have a wider range of experience than specialists and thus are less likely to have developed special bureaucratic personalities. Also, as consistent with my secondary theme in earlier examples, nonspecialists by nature are less isolated than specialists. The use of nonspecialists may break the isolation necessary for the development of dysfunctional bureaucratic personalities.

Economic models of bureaucracy have typically been based on principal-agent theory. Their purpose is to derive optimal organizational structures, contingent on the technical nature of information flows. (Two excellent examples are Paul Milgrom and John Roberts, 1988, and Bengt Holmstrom and Jean Tirole, 1988). In contrast, my analysis suggests an alternative way in which information affects the performance of a bureaucracy. Bureaucratic structures that make specialized decisions may behave in "deviant" ways. In special cases such as dedicated scientists in the laboratory, the Green Berets, or the U.S. Forestry Service (see Herbert Kaufman, 1960), this isolation may be beneficial and the deviance quite functional.¹³ On the other hand, as we have

¹³In the case of scientists, it may not just be individual but also group psychology under isolation that results in the odd scientific personalities exemplified by the "mad scientist" image.

seen, this same specialization may be dysfunctional. Entirely absent from the principal-agent model is the possibility that behavior changes occur latently in response to obedience to authority. While the theory of bureaucracy must address incentive problems (as in principal-agent problems), it should also consider the need to organize decision making so as to create functional (rather than dysfunctional) changes in personalities.

IX. Conclusion

Standard economic analysis is based upon the Benthamite view that individuals have fixed utilities which do not change. Stigler-Becker and Becker-Murphy have gone so far as to posit that these utilities do change, but that individuals are forward looking and thus foresee the changes that will occur. A more modern view of behavior, based on twentieth-century anthropology, psychology, and sociology is that individuals have utilities that do change and, in addition, they fail fully to foresee those changes or even recognize that they have occurred. This lecture has modeled such behavior in sequences of decisions, given examples from everyday life, indicated the situations in which such behavior is likely to occur, and, in some instances, suggested possible remedies. The theory of procrastination and obedience has applications to savings, crime, substance abuse, politics, and bureaucratic organizations.

REFERENCES

- Aronson, Elliot, *The Social Animal*, 4th ed., New York: Freeman, 1984.
- Arrow, Kenneth J., *Social Choice and Individualistic Values*, 2nd ed., New Haven: Yale University Press, 1963.
- Asch, Solomon E., "Effects of Group Pressure upon the Modification and Distortion of Judgments" in Harold S. Guetzkow, ed., *Groups, Leadership and Men*, Pittsburgh: Carnegie Press, 1951.
- _____, *Social Psychology*, Englewood Cliffs: Prentice-Hall, 1952.
- Becker, Gary S., "Crime and Punishment," *Journal of Political Economy*, March/April 1968, 76, 169-217.
- _____, and Murphy, Kevin M., "A Theory of Rational Addiction," *Journal of Political Economy*, August 1988, 96, 675-700.
- Borgida, Eugene and Nisbett, Richard, "The Differential Impact of Abstract vs. Concrete Information on Decision," *Journal of Applied Social Psychology*, July 1977, 7, 258-71.
- Brown, Roger, *Social Psychology: The Second Edition*, New York: Macmillan, 1986.
- Burka, Jane B. and Yuen, Lenora M., *Procrastination: Why You Do It, What to do About It*, Reading: Addison-Wesley, 1983.
- Cagan, Phillip, *The Effect of Pension Plans on Aggregate Saving: Evidence from a Sample Survey*, NBER, Occasional Paper No. 95, New York: Columbia University Press, 1965.
- Carpenter, Cheryl et al., *Kids, Drugs and Crime*, Lexington: D.C. Heath, 1988.
- Dahlin, Michel R., "From Poorhouse to Pension: The Changing View of Old Age in America, 1890-1929," unpublished doctoral dissertation, Stanford University, 1983.
- Deutscher, Isaac, *Stalin: A Political Biography*, Oxford: Oxford University Press, 1949.
- Ellwood, David T. and Summers, Lawrence H., "Poverty in America: Is Welfare the Answer or the Problem?," in Sheldon H. Danziger and Daniel H. Weinberg, eds., *Fighting Poverty: What Works and What Doesn't*, Cambridge: Harvard University Press, 1986.
- Feldman, Harvey W., Agar, Michael H. and Beschner, George M., *Angel Dust: An Ethnographic Study of PCP Users*, Lexington: D. C. Heath, 1979.
- Fisher, Irving, *The Theory of Interest*, New York: Macmillan, 1930.
- Galanter, Marc, "Psychological Induction Into the Large-Group: Findings from a Modern Religious Sect," *American Journal of Psychiatry*, December 1980, 137, 1574-1579.
- _____, *Cults: Faith, Healing, and Coercion*, New York: Oxford University Press, 1989.
- _____, et al., "The 'Moonies': A Psychological Study of Conversion and Membership in a Contemporary Religious Sect,"

- American Journal of Psychiatry*, February 1979, 136, 165-70.
- Gerstel, David U., *Paradise, Incorporated: Synanon*, Novato: Presidio Press, 1982.
- Hirschman, Albert O., *Exit, Voice and Loyalty: Responses to Decline in Firms, Organizations and States*, Cambridge: Harvard University Press, 1970.
- Holmstrom, Bengt R. and Tirole, Jean, "The Theory of the Firm," in Richard Schmalensee and Robert Willig, eds., *The Handbook of Industrial Organization*, Amsterdam: North-Holland, 1988.
- Hurd, Michael D., "Research on the Elderly: Economic Status, Retirement, Consumption, and Saving," *Journal of Economic Literature*, June 1990, 28, 565-637.
- Janis, Irving L., *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*, Boston: Houghton Mifflin, 1972.
- Katona, George, *Private Pensions and Individual Savings*, Monograph No. 40, SRC, Institute for Social Research, Ann Arbor: University of Michigan, 1965.
- Kaufman, Herbert, *The Forest Ranger: A Study in Administrative Behavior*, Baltimore: Johns Hopkins University Press, 1960.
- Khrushchev, Nikita S., *Khrushchev Remembers: The Glasnost Tapes*, transl./ed., Jerrold L. Schector with V. V. Luchkov, Boston: Little, Brown, 1990.
- MacLeod, Jay, *Ain't No Makin' It*, Boulder: Westview, 1988.
- Matza, David, *Becoming Deviant*, Englewood Cliffs: Prentice-Hall, 1969.
- _____, *Delinquency and Drift*, New York: Wiley & Sons, 1964.
- Merton, Robert K., *Social Theory and Social Structure*, 1968 Enlarged Edition, New York: Free Press, 1968.
- Milgram, Stanley, *Obedience to Authority: An Experimental View*, New York: Harper and Row, 1975.
- _____, "Some Conditions of Obedience and Disobedience to Authority," in I. D. Steiner and M. Fishbein, eds., *Current Studies in Social Psychology*, New York: Holt, Rinehart, and Winston, 1965.
- Milgrom, Paul R. and Roberts, John, "An Economic Approach to Influence Activities in Organizations," *American Journal of Sociology*, Suppl. 1988, 94, S154-S179.
- Munnell, Alicia H., "Private Pensions and Savings: New Evidence," *Journal of Political Economy*, October 1976, 84, 1013-32.
- Nisbett, Richard E. and Ross, Lee, *Human Inference: Strategies and Shortcomings of Social Judgment*, Englewood Cliffs: Prentice-Hall, 1980.
- Ofshe, Richard, "The Social Development of The Synanon Cult: The Managerial Strategy of Organizational Transformation," *Sociological Analysis*, Summer 1980, 41, 109-27.
- Ouchi, William G., *Theory Z: How American Business Can Meet the Japanese Challenge*, Reading: Addison-Wesley, 1981.
- Peters, Thomas J. and Waterman Jr., Robert H., *In Search of Excellence: Lessons from America's Best-Run Companies*, New York: Harper and Row, 1982.
- Ross, Jerry and Staw, Barry M., "Expo 86: An Escalation Prototype," *Administrative Science Quarterly*, June 1986, 31, 274-97.
- Ross, Lee, "Review of Arthur G. Miller, *The Obedience Experiments: A Case Study of Controversy in Social Sciences*, New York: Praeger, 1956," *Contemporary Psychology*, February 1988, 33, 101-04.
- Schelling, Thomas C., "Enforcing Rules on Oneself," *Journal of Law, Economics and Organization*, Fall 1985, 1, 357-73.
- Schorr, Lisbeth B., *Within Our Reach*, New York: Doubleday, 1989.
- Staw, Barry M., "Knee-Deep in the Big Muddy: A Study of Escalating Commitment to a Chosen Course of Action," *Organizational Behavior and Human Performance*, June 1976, 16, 27-44.
- _____, and McClane, S., *Throwing Good Money After Bad: Escalation in a Banking Context*, Symposium presentation, Academy of Management, 44th Annual Meeting, Boston, 1984.
- _____, and Ross, Jerry, "Behavior in Escalation Situations: Antecedents, Prototypes, and Solutions," in Larry L. Cummings and Barry M. Staw eds., *Research in Organizational Behavior*, Vol. 9, Greenwich: JAI Press, 1987.
- Stigler, George J. and Becker, Gary S., "De Gustibus Non Est Disputandum," *Ameri-*

- can Economic Review*, March 1977, 67, 76-90.
- Stotland, E. and Canon, Lance K.**, *Social Psychology: A Cognitive Approach*, Philadelphia: Saunders, 1972.
- Thaler, Richard H. and Shefrin, Hersh, M.**, "An Economic Theory of Self-Control," *Journal of Political Economy*, April 1981, 89, 392-406.
- Venti, Steven F. and Wise, David A.**, "IRAs and Savings," NBER Working Paper No. 1879, Cambridge, April 1986.
- _____ and _____, "Have IRAs Increased U.S. Saving?: Evidence from Consumer Expenditure Surveys," NBER Working Paper No. 2217, Cambridge, April 1987.
- Whyte, William F.**, *Street Corner Society*, Chicago: University of Chicago Press, 1943.
- Wilson, William J.**, *The Truly Disadvantaged: The Inner City, the Underclass and Public Policy*, Chicago: University of Chicago Press, 1987.
- U.S. Bureau of the Census**, *Statistical Abstract of the United States: 1987*, Washington: USGPO, 1987.

The Economics Major: Can and Should We Do Better than a B—?

By JOHN J. SIEGFRIED, ROBIN L. BARTLETT, W. LEE HANSEN,
ALLEN C. KELLEY, DONALD N. MCCLOSKEY,
AND THOMAS H. TIETENBERG*

Yes, and yes.

The two questions asked in the title of this paper were prompted by an invitation to the American Economic Association's Committee on Economic Education to participate with eleven other disciplines in a national review of arts and sciences majors initiated by the Association of American Colleges (AAC). The goal of this "Project on Study-in-Depth" was to evaluate the economics major within the liberal arts curriculum, highlighting connections and interactions among disciplines. In the process we examined the economics major in considerable detail. The result was a lengthy report,¹ from which the present paper abstracts key findings, focusing on the purpose of the major, recommendations for improvement, and methods for effecting change.

In compiling the AAC report we became painfully aware that while our collective experience in the field of "economics education" provides us with considerable background, the representativeness of our

judgments required checking. To this end, seminars were given at several colleges and universities and over 100 copies of the report were circulated to other economists, including leading economic educators; feedback was incorporated in subsequent versions. In addition, in May 1990 a survey of economics faculty at 127 colleges and universities was undertaken with the goal of vetting and evaluating our main conclusions and recommendations.

As a part of that survey, respondents were invited to complete a "report card" (A,...,F) based on twelve criteria evaluating the effectiveness of the economics major at their institutions.² Tabulated on a 4.0 grading scale, the economics major earned an overall grade of B—. An analysis of the detailed statistical tabulations and open-ended responses to some thirty-seven questions suggested how it might be improved.

In this paper we argue that the quality of the major is suboptimal, a situation probably resulting from two decades of expanding enrollments and the relative popularity of the major, which occurred just when demands on faculty for research and other responsibilities greatly increased, and when college administrations tightened con-

[†]*Discussants:* David Colander, Middlebury College; Claudia Goldin, Harvard University; Alan Blinder, Princeton University; Eric Hanushek, University of Rochester.

*Professors of Economics at Vanderbilt University, Denison University, University of Wisconsin-Madison, Duke University, University of Iowa, and Colby College, respectively. Funding for this project was provided by grants from the Fund for the Improvement of Postsecondary Education and the Ford Foundation to the Association of American Colleges.

¹The Economics Major in American Higher Education," *Journal of Economic Education*, No. 3, Summer 1991, 22. The present paper defers to that article for all literature references.

²The criteria included aspects of course content, the institutional environment (for example, manageability of class size, effectiveness of instructional methods, adaptation to student learning styles, connections with other courses in economics), and skills developed (analytical, mathematical, empirical, critical judgment, applications, oral/written communication, ability to think like an economist). A copy of the questionnaire, list of participating schools, and survey results is available from the Joint Council on Economic Education, 432 Park Avenue South, New York, 10016.

straints on faculty size in preparation for expected declines in enrollments in the 1980's. This stocktaking indicates that the economics major merits improvement. Here we provide specific suggestions to accomplish that objective.

I. The Goal of the Major

Enabling students to develop a capacity to "think like an economist" is the overarching goal of the major. All other virtues follow. Thinking like an economist involves using chains of deductive reasoning in conjunction with simplified models (such as supply and demand, benefit-cost analysis, and comparative advantage) to illuminate economic phenomena. To some, economists tend to abstract too much from the richness of human behavior and reality; to many economists, the strength of our analysis is the provision of focus and clarity of thought; parsimonious models are a virtue, not a vice.

Thinking like an economist also involves identifying and evaluating tradeoffs in the context of constraints, distinguishing positive from normative analysis, and tracing behavioral implications of change while abstracting from aspects of reality. It, moreover, involves describing redistributive implications of change, amassing data to evaluate economic events, and testing hypotheses about how consumers and producers make choices and how the economy works. Finally, thinking like an economist involves examining many problems through a filter of efficiency—coping with limited resources.

Thinking like an economist requires creative skills, too. Identifying economic issues and problems, framing them in ways other people do not see, devising novel policy proposals for dealing with problems, analyzing both the intended and unintended effects of policies, and devising innovative methods to estimate the magnitude of these effects—all are as central to the discipline as is the development of logically coherent theories.

This statement of purpose appeals to the vast majority of economics faculty in American colleges and universities. But how can

we be assured that our undergraduate students really meet this objective by the time they walk across the graduation platform? Here we present a blueprint for reform of the economics curriculum which we believe will help students better understand how to think like an economist.

II. Recommendations for Improving the Major

An ideal program for the major includes *foundation requirements* (introductory macro and micro, intermediate macro and micro theory, and quantitative methods), a *breadth requirement* (elective courses), a *depth dimension* which probes more deeply the structure of knowledge in elective courses, and a *capstone experience*, whereby students apply their knowledge and skills in creative and systematic ways through research and writing.

We believe substantial opportunity exists for improving each component of the curriculum. In this section we identify areas of deficiency and make recommendations for change.

A. Foundations I: The Introductory Courses

Introductory courses expose students to economic theory, principles, and concepts, and instruct them how to apply the economists' tools to analyze problems and unfamiliar situations. These courses typically serve both economics and business majors and students attempting to satisfy the institution's general education requirements.

Deficiencies. Introductory courses tend to be encyclopedic, and all too often oriented toward formalism of theory at the expense of application.

Recommendation.

1) Emphasize the *application* of a limited number of important concepts and theoretical tools to a variety of problems, at the expense of some of the existing formal and detailed elaboration of theoretical constructs or the extent of coverage of the vast array of topics contained in most textbooks.

Our survey indicates that this orientation and recommendation commands considerable support. In fact, it represents the dominant approach of most existing intro-

ductory economics courses. Fortunately, these courses constitute a relatively sound component of the major at most colleges and universities.

B. *Foundations II: The Intermediate Theory Courses*

Intermediate theory courses accomplish three goals. First, they show how economists develop and use theory, and how rigorous thinking illuminates economic phenomena. Second, they provide prerequisite tools required to undertake economic analyses in elective courses. Third, they provide important information or signals to students: what the major is like, what content must be mastered, what skills must be developed, and what standards of performance must be met.

Deficiencies. While most intermediate macro and micro theory courses develop well the rigor and elegance of economic theory, they tend to slight a critically important issue: how theory and real world events interact to produce new knowledge, concepts and theories, and how such theories are evaluated. In particular, the "usefulness" of theoretical topics and paradigms, largely evaluated by confronting theory with data; applying models to various problems, and comparing the outcomes of alternative theoretical constructs, merits greater emphasis.

Recommendations.

- 1) Coordinate content so that these courses establish a foundation of knowledge and skills on which other courses and instructors can rely.
- 2) Establish explicit connections between theory and its empirical counterparts, to help students appraise the importance of theoretical constructs, to provide a basis for selecting assumptions, and to show that theory is relevant.

Again, students are reasonably well served by these courses, although our second recommendation merits emphasis. Use and appraisal of theory should in practice constitute the primary basis for establishing the course "standard"—the level of difficulty, as it were, for screening students interested

in the major. While facility with formal theoretical constructs should represent a necessary condition for passing these courses, it should not be sufficient. Students must, in addition, be able to *use and appraise* theory—a challenging, but arguably a central skill as a "gateway to the major."

C. *Foundations III: The Quantitative Methods Courses*

Quantitative methods courses expose students to the empirical dimensions of economics: the techniques for testing and evaluating our understanding of economic phenomena and economic behavior. These courses usually emphasize algebraic skills and statistical techniques.

Deficiencies. The dominant attention to the formal aspects of statistical analysis shortchanges students' exposure to how these methods are applied in the empirical work economists do, and to the criteria for and methods of selecting and appraising the data used in empirical analysis.

Recommendations.

- 1) Reorient the course from its almost singular statistical orientation to emphasize a wider range of quantitative methods pertinent to economic analysis.
- 2) Increase the allocation of time to the most widely used (but still accessible) methods employed by economists in their empirical research.
- 3) Devote more attention to limitations imposed by the absence of appropriate data, poor quality of existing data, and use of proxy variables; illustrate creative attempts to overcome these difficulties.

We propose a broader quantitative methods course, one that devotes more attention to research design and to the context within which quantitative methods are applied. The former requires giving students a better sense of what is involved in research—from formulating the underlying question to conceptualizing the study and organizing the process for completing it. The latter requires knowing something about measuring economic behavior and phenomena; selecting, organizing, appraising, and manipulating economic data; having the ability to test

hypotheses; and knowing how to interpret the results of various statistical procedures.

D. Breadth

Breadth is typically achieved by students self-selecting elective courses to explore subfields and topics within economics, which draw upon the foundation courses, and are often enriched by historical, institutional, and empirical detail. As few as three, and as many as six or seven courses are taken, and the choice is usually unconstrained.

Deficiencies. This format all too often results in students acquiring a narrow, parochial perspective, unable to come to grips with deviations from marginalist thinking, and incapable of dealing sensibly with problems that involve approaches different from atomistic models of individual choice.

Recommendations.

Require at least four or five elective courses.

Structure student choices to produce greater breadth: we recommend at least one course each in 1) contextual, 2) international, and 3) public sector economics.

Contextual inquiry includes courses in economic history (where connections between economics and history are explicit), history of economic thought (where different modes of thought are exposed), comparative economic systems (where social/political/cultural dimensions that influence distinctive economic systems are compared), and area studies (where synthetic analyses of countries and regions are explored). Such courses take the edge off narrow thinking about economics, and they illuminate the importance of context and structure (initial conditions and constraints) that shape the dominant "marginalist" orientation of economics.

International courses include not only traditional offerings in trade and finance, but also those in economic development, comparative systems, and traditional fields that may incorporate a notable international component (for example, the multinational corporation). Such courses expand students' perspective from the parochial to the global, placing them in a stronger posi-

tion to use their tools of economic inquiry in a world that is rapidly becoming more integrated.

Public sector economics courses include not only the usual ones in public finance and taxation, but also some offerings in theory (stressing public goods, externalities, collective decision making, and market failure), labor economics (stressing aspects of labor regulation), and the like. Such courses simultaneously illuminate and qualify the role of individual, free market choice, a dominant paradigm in economics. Students should gain considerable understanding of the applicability, methods, and limitations of collective choice, including nonmarket options for resource allocation. These dimensions of decision making account for one-third to almost all resource allocations in most countries, and they are too important to relegate to a few weeks of exploration in the foundation courses.

Feedback from our survey about the breadth recommendations was strongly positive, although some respondents expressed legitimate concern about the administrative feasibility of offering enough courses on an annual basis to provide such breadth. This concern can be addressed by defining the breadth options in ways that are consistent with the capacity of institutions to staff such courses.

E. Depth

Depth in elective courses is accomplished by going beyond the coverage of textbook-oriented field courses to how current knowledge evolved and how new knowledge is developed in the field.

Deficiencies. While students often leave elective courses with sound "state of the art" surveys of coverage, they often lack a feeling for the field's central research issues, how knowledge about them has developed over time, and how the ideas revealed in the courses are related to the fundamental theories of economics.

Recommendations.

1) Link the important ideas in elective courses to general principles introduced in the theory courses; at least some of the

elective courses should require an intermediate theory course as a prerequisite.

2) Examine how economic phenomena and behavior are observed and measured, pertinent empirical data are appraised and selected to illuminate these phenomena, and how these data can be effectively organized and analyzed.

3) Explore the process of expanding existing knowledge and developing new knowledge in the field.

F. Capstone Experience

A capstone experience can help complete the process of intellectual maturation. Such an experience is achieved by giving students opportunities to apply what they have learned to an economic question or problem—in effect, “doing economics.” In the process they acquire an increased capacity to “think like an economist.” Completion of such an experience is typically accomplished by writing a senior thesis, carrying out an honors research project, or enrolling in an independent study course. In some institutions, special senior seminars facilitate these opportunities.

Deficiency. Too few institutions offer these opportunities.

Recommendation.

1) Provide more opportunities for advanced students (seniors) to “do economics” by undertaking a substantial independent project that requires majors to formulate an underlying question or problem, to structure an analysis, to assemble the information necessary to carry out the analysis, to draw conclusions from the results, and to communicate the findings to others in both oral and written form.

While such a capstone experience is educationally sound for all majors, in practice departmental resources may be constrained. When constraints prevent extending this opportunity to all majors, such courses are most productively offered to more capable students, often in an honors program.

III. Getting It Done

A respectable economics major that teaches students how to think like an

economist requires considerable instructional resources, especially if, as we argue, students must obtain extensive practice at really *doing* economics. To be successful, this requires relatively small classes, 20–25 students in intermediate macro and micro and elective courses, and approximately 15 students in courses emphasizing writing, oral presentations and argumentation, and research projects.

Deans and chairs will immediately observe that such a major is expensive, and thus “compromises” must be made. Counterarguments are that economics carries more than its share of costs with its large introductory courses; indeed, its cost per credit hour is likely to be lower than those of many other departments and much lower than those of many science courses with their laboratory and discussion sections. More important, the low-cost technology of large classes, the lecture format, and multiple-choice examinations (so prevalent throughout the economics curriculum) have often resulted in majors simply being “exposed” to economics in varying degrees. Lamentably, in all too many instances, even the *minimum* mastery level of understanding how to think like an economist is sacrificed.

When the production of minimally acceptable output requires a much higher-cost technology, how is it possible to make the major work? The answer, it seems to us, is simple: either increase faculty or ration access to the major to fit the resources available while maintaining quality standards and fulfilling the responsibilities of each college or university. Placing a limit on the number of economics majors will conflict with the “philosophy” of many institutions. However, surely unconstrained access to the major without concomitant resources, resulting in diminished standards that compromise the intellectual integrity of the enterprise, is also at variance with prevailing educational philosophy. Responsible educational planning requires “living within one’s budget” of instructional resources. Thus, the question of *how* to ration access to the major becomes paramount.

The method of rationing may vary from school to school, depending on the institu-

tion's policies and procedures. Whatever method is used, however, it should be *educationally sound* with respect to the *goals* of the major. Our own preference is to offer intellectually challenging intermediate micro and macro and quantitative methods courses whose "reputation" ensures that the number of students intending to major does not exceed capacity.

What does and does not constitute "intellectual challenge" in such courses must be spelled out. It does *not* require the use of formal (and seemingly difficult or sophisticated) tools (mainly mathematics); and it does *not* involve the use of unfair or tough grading standards, unreasonable assignments, or "scare tactics" as techniques to discourage enrollments. It *does* require holding students to the standard of properly *applying* reasonably sophisticated economic ideas to a variety of unfamiliar problems. This standard is intellectually more demanding than facility with formal tools per se, and it is, in fact, the best early indicator of whether a student has the ability to come to grips with the major. (Parenthetically, introductory courses should *not* be used to ration access to the major since such courses should be widely

accessible to nonmajors and students of diverse backgrounds and goals.)

As a result, students should be in classes sufficiently small to permit them to interact effectively within their instructors. Professors should then be expected to employ evaluation methods that give students an opportunity to develop and use writing and oral skills. Learning should take a more active form, and therefore have longer lasting effects.

The undergraduate economics major has slipped in quality over the past two decades as large enrollments undermined standards. We see no reason why large *enrollments* in economics programs need pose a problem. Indeed, offering exceptionally high-quality introductory economics courses (even if taught in large classes) should be a primary goal of economics departments. A related goal is to ensure that economics is one of the most exciting and intellectually challenging majors. Having said all this, we believe the central task is to make certain that economics majors understand how to "think like an economist"—surely a highly demanding but attainable goal. To accomplish this, tough choices—the hallmark of economics—must be made.

An Agenda for Research on Economic Education in Colleges and Universities

By WILLIAM BECKER, ROBERT HIGHSMITH, PETER KENNEDY,
AND WILLIAM WALSTAD*

The quantity of research on economic education at the college and university level declined during the past decade. In the 1980-90 period, the number of research-related articles on economics instruction in higher education fell by about 17 percent from the number published during the 1969-79 period. A possible reason for this reduction may have been the publication of a review of research on economic education at the college and university level by John Siegfried and Rendigs Fels (1979). This extensive survey may have inadvertently led researchers to believe that most of the major topics at this level had been studied and that further research would not yield insights. Another reason could have been the success of the Joint Council on Economic Education in directing resources to precollege issues.

Whatever the reasons for the college-level decline, it is disturbing because we think the teaching of economics in colleges and universities can be improved by research on what influences the delivery and the effectiveness of instruction. In our view, research on economic education at the postsecondary level should be directed to three major areas. First, the multiple outputs from learning economics need to be defined, measured, and investigated so that a fuller range of benefits from studying economics can be incorporated into decisions about courses and degree programs. Second, more emphasis should be placed on the analysis

of the economics major, as distinct from individual courses, to enhance the structuring of programs. Third, the replication of earlier research is required to determine the extent to which conclusions drawn from those studies still hold and to relate those findings to new developments.

I. Multiple Outputs

College economics courses may contribute more to student development than can be measured by scores on a cognitive test. W. Lee Hansen (1986), for example, described five "proficiencies" that he thinks students should be able to demonstrate from majoring in economics (i.e., gaining access to existing knowledge, displaying command of existing knowledge, displaying the ability to draw out existing knowledge, utilizing existing knowledge to explore issues, and creating new knowledge). This demonstration would be difficult in a multiple-choice testing framework.

Recognition of the multiple outputs from studying economics has not gone unnoticed in the research literature in economic education. Hansen, Allen Kelley, and Burton Weisbrod (1970) discussed the need for more research on how desired outcomes and how the distribution of benefits from teaching differ among students. Judith Yates (1978) complained that our research horizons were too narrow because of our focus on easily measured outputs. Eric Hanushek (1979; 1986) described the importance of joint products in the educational process, and, in particular, faulted researchers in economic education for their singular emphasis on multiple-choice tests. Despite the widespread knowledge that there are multiple outputs from teaching, few studies in economic education have incorporated this fact.

*Indiana University, Bloomington, IN 47405; Joint Council on Economic Education, New York, NY 10016; Simon Fraser University, Burnaby, B.C. Canada V5A 1S6; and University of Nebraska, Lincoln, NE 68588-0402, respectively. This paper is an abridged version of a forthcoming *Journal of Economic Education* (Summer 1991) article. Constructive criticism on earlier drafts was provided by W. Lee Hansen, John Siegfried, and Michael Watts.

The problem of deciding how to value outcomes also must be considered. For example, William Becker, William Greene, and Sherwin Rosen (1990) argue against the use of change-score models in economic education research because these models ignore the fact that the market for new graduates does not place a value on student learning as much as it values the final level of accomplishment. They assert that the normative beliefs of an instructor or entire faculty about the importance of given intellectual skills is elusive without reference to what employers are paying for the bundle of skills embodied in the college graduate, and what they desire from the graduate. Becker and William Walstad (1990) also show that data loss from pretest to posttest may pose problems for assessing value added.

Debates on problems associated with the value added by education make clear that research on a wider coverage of outcomes from instruction will not be easy to conduct. Each output must first be clearly identified and accurately measured. A consensus also needs to be reached among researchers about what is required to assess each outcome. The scope of research work then must be broadened to include an array of outcomes, if they can be measured and are considered to be important.

Several research topics to be addressed in the context of multiple outputs seem to have particular relevance in considering inputs as well. First, what is the role of basic skills (i.e., reading, writing, computing, and mathematics) in economics courses? For example, do students who are asked to use mathematics in their courses gain a better understanding of economics than students who use limited mathematics? Conversely, does the study of economics significantly improve mathematics skills? It seems reasonable to think that basic mathematics skills are both an input for economics learning and an output from economics learning, but at present we have limited empirical data on these possible relationships. Better knowledge of the interaction between basic skills as inputs and outputs from economics instruction would be valuable for identifying prerequisites for courses.

A second subject for study is the relative merits of fixed-response (multiple choice) and constructed-response (essay or short answer) tests for measuring student achievement in economics. Although the shortcomings of fixed-response tests are well known, the benefits of reliable and valid measures of a multiple-choice-type test for evaluation and research may not be outweighed by the negatives. What is needed is further exploration of the relationship between student performance on fixed-response vs. constructed-response tests. Do the different tests capture different dimensions of student performance, or are they measuring essentially the same dimension? The value of fixed-response vs. constructed-response measures is debated among faculty members. There is a literature on the topic in other fields (see Hunter Breland et al., 1987), but the issue has not been thoroughly investigated in economics.

A third topic is research on learning retention. Research typically measures cognitive outcomes during an economics course. With the notable exception of Phillip Saunders (1980), no extensive research exists on the lasting effects from economics coursework because longitudinal data typically are not available. These data should be collected to investigate what students retain *X* years after completing coursework and a major in economics. Presumably there is a host of factors that explain retention from a course, a series of courses, or from the major. We also suspect that the level of retention differs across the multiple outputs.

A fourth area for research is the effect of instructors on student outcomes. The target for the work would be the identification of the attributes that exemplary teachers of economics develop as they accumulate more classroom experience; knowledge of the characteristics of good teaching should enable new instructors to shorten the time it takes to become better teachers. We also suspect that exemplary teachers, consciously or unconsciously, strongly affect more than one dimension of student development. A starting point for this research would be to survey economists who have won awards for

excellence in teaching. These findings could be compared to results from similar surveys in other disciplines, or to the extensive research on student evaluation of instruction in economics. Psychologists have also studied differences in the characteristics of "expert" and "novice" approaches to problem solving, and research of this type might have application to economics teaching.

II. The Economics Major

Although a few studies (for example, Siegfried and Jennie Raymond, 1984), have examined the economics major, more can be learned about the demand curve for course enrollments and majors. One question yet to be answered, for example, is why students take courses and/or major in economics. A related query is why students elect certain courses in economics when majoring in the subject. Factors such as course difficulty, instructor or department reputation, recommendations of peers, preparation for graduate school, or personal characteristics might be included in a list of factors influencing course selection or the decision to major in economics.

There are many alternative routes to a major in economics; for example, majors may have a business or social science orientation with emphasis on either applied or theoretical work. We need to know the benefits and drawbacks of these alternatives. In addition, the claim is often made that training in economics is central to an understanding of the complexities of the world. It is not clear to what extent this claim is justified relative to those of other social sciences or areas of study. How do economics majors evaluate their educational experience compared with students majoring in business or other social sciences?

Addressing these concerns will initially involve some survey work. The questions we are posing, however, are broader and focus on assessing the relative value of an education in economics as perceived by students who are taking classes or by those who have completed a course of study. Expansion of the data sources will be necessary for some of the evaluations. For example, alumni

records from colleges and universities might be exploited for examining the performance of students after graduation or for evaluating performance differences across alternative economics majors. Moreover, "successful persons" (top corporate executives, government officials, or community leaders) who have majored in economics could be contacted and asked how majoring in economics contributed to their success. Despite the potential for sample selection problems, a follow-up survey of alumni or the most successful alumni might identify features of economics that should be emphasized by departments.

III. Replications

Fundamental to sound research is the ability to replicate results. Economic education is sufficiently mature as a research area that some of the "accepted" results should be reexamined. Replication should not be mere duplication; it should also provide opportunities to extend previous work in new directions.

The questions that should be reexamined are numerous. One topic studied in the 1970's was how high school preparation in economics affects performance in the college principles of economics classes or the decision to select economics as a major (see Saunders, 1970). Since those studies were conducted, many changes have occurred in high school courses. A variety of "economics" courses are being mandated by states, and there is now an Advanced Placement course in economics (Stephen Buckles and John Morton, 1988). There are also more instructional materials available, and teachers may be better prepared to teach economics. These changes suggest that college students with experience in high school economics may have advantages that bear on their placement and treatment in college-level economics courses.

A second topic is the effects of class size. Although students tend to dislike large classes, many studies of introductory courses have found that class size has little influence on multiple-choice test scores once class size rises above a threshold level (David

Williams et al., 1985). Henry Raimondo, Louis Esposito, and Irving Gershenberg (1990), however, found that large classes at the intermediate level can be detrimental in some areas of economics. The work of David Card and Alan Krueger (1990) suggests that class size, as well as other expenditure variables, may affect the financial returns from education. The critical role of class size on the multiple outputs of economic education needs to be established.

A third topic, often debated in economics departments, is the optimum order in which courses are taken. For instance, does it matter if students take the microeconomics course before the macroeconomics course? Possibly because of differences in output measures and research design, studies by John Fizel and Jerry Johnson (1986) and James McCoy, David Brasfield, and Martin Milkman (1989) suggest conflicting answers to such questions. Related to these replications would be studies that address possible efficiency gains from condensing the two-semester principles courses into a one-semester course.

A fourth topic, controversial in many larger economics departments, is the increasing employment of graduate student instructors for whom English is a second language. Complaints have been raised by students, parents, and legislators about whether undergraduate students are able to perform as well with nonnative English-speaking instructors as they are able to do with native English-speaking instructors. Michael Watts and Gerald Lynch (1989) found that students with instructors for whom English was a second language performed less well than students with instructors for whom English was a first language. If this finding is supported by studies at other institutions, steps may be needed to upgrade the language and teaching skills of these potential instructors before they are permitted to teach.

A fifth topic is how student and instructor differences affect student outcomes and instructor performance. Studies have been done on the relationship between teaching style and learning styles. Replication of previous work and further study along these

lines may identify ways the teaching of economics can be structured to achieve a better match between the teaching styles of instructors and learning styles of students. Also, we need to know more about the role of gender, race, ethnic background, socioeconomic status, and general ability because they are likely to affect the multiple outcomes from economics instruction and related issues such as the decision to major in economics.

A sixth topic is the role of new technologies. In the 1960's, televised lectures were touted as a means to meet college enrollment demands. Mainframe computers were considered in the 1970's to add variety to classroom lectures and to manage instruction. The 1980's witnessed the introduction of the microcomputer and improvement in videotape technology. Yet few of these innovations got beyond the experimental stage. The applicability of research on past innovations to current innovations is of questionable value. We know little about the cost effectiveness of newer technologies on student performance in a course or in the major.

IV. Concluding Comments

We urge researchers in economic education to mobilize in this decade, as they did in the 1970's, to advance our understanding of the teaching of economics at the postsecondary level. The quality of work in economic education, at all levels, is steadily improving and researchers are finding more imaginative ways of addressing problems both old and new. Although by publishing this agenda we hope to influence the direction of future research in this area, we hope also that researchers will feel free to attack the problems they perceive to be important, in the ways they see fitting.

We recognize that some of our current agenda, in particular the part relating to multiple outputs, involves work requiring considerable resources and the coordination/interaction of several researchers. The AEA Committee on Economic Education (CEE) and the Joint Council on Economic Education (JCEE) must be encouraged to

develop a "social infrastructure" that allows individual researchers to tackle this agenda in an efficient manner. In this regard, we offer two final recommendations.

If multiple outputs are to play a prominent role in future research, individual researchers should be given guidance on what outputs are thought to be relevant, and how they are to be measured. We recommend that the JCEE, in cooperation with the CEE, solicit position papers on output measures from a range of scholars in a variety of disciplines, convene a conference to debate this issue, obtain a consensus on the outputs that should be used for assessment purposes, and develop instrumentation and data bases for these new output measures.

Much of the agenda overlaps extensively with research in education and in other disciplines. Economics is not the only discipline with an "education" division, and many learning concepts described in the education literature could provide fertile ground for research in economic education. Researchers in economic education should become more conversant with this education literature. To facilitate this, we recommend that the JCEE and the CEE commission surveys of relevant dimensions of this literature, with particular emphasis on how results in this literature relate to results in the economic education literature.

REFERENCES

- Becker, William and Walstad, William, "Data Loss from Pretest to Posttest As a Sample Selection Problem," *Review of Economics and Statistics*, February 1990, 72, 184-88.
- _____, Greene, William and Rosen, Sherwin, "Research on High School Economic Education," *American Economic Review Proceedings*, May 1990, 80, 14-22.
- Breland, Hunter et al., *Assessing Writing Skills*, New York: College Entrance Examination Board, 1987.
- Buckles, Stephen and Morton, John, "The Effects of Advanced Placement on College Introductory Economics Courses," *American Economic Review Proceedings*, May 1988, 78, 263-68.
- Card, David and Krueger, Alan, "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," NBER Working Paper No. 3358, 1990.
- Fizel, John and Johnson, Jerry, "The Effects of Macro/Micro Course Sequencing on Learning and Attitudes in Principles of Economics," *Journal of Economic Education*, Spring 1986, 17, 87-98.
- Hansen, W. Lee, "What Knowledge Is Most Worth Knowing—For Economics Majors," *American Economic Review Proceedings*, May 1986, 76, 149-52.
- _____, Kelley, Allen C. and Weisbrod, Burton, "Economic Efficiency and the Distribution of Benefits from College Instruction," *American Economic Review Proceedings*, May 1970, 60, 364-69.
- Hanushek, Eric A., "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," *Journal of Human Resources*, Summer 1979, 14, 351-88.
- _____, "The Economics of Schooling," *Journal of Economic Literature*, September 1986, 24, 1141-77.
- McCoy, James, Brasfield, David and Milkman, Martin, "Course, Sequence and Student Performance in Principles of Economics," paper presented at the Southern Economics Association Conference, November 1989.
- Raimondo, Henry, Esposito, Louis and Gershenson, Irving, "Introductory Class Size and Student Performance in Intermediate Theory Courses," *Journal of Economic Education*, Fall 1990, 21, 369-81.
- Saunders, Phillip, "Does High School Economics Have A Lasting Impact?," *Journal of Economic Education*, Fall 1970, 1, 39-55.
- _____, "The Lasting Effects of Introductory Economics Courses," *Journal of Economic Education*, Winter 1980, 12, 1-14.
- Siegfried, John and Fels, Rendigs, "Research on Teaching College Economics: A Survey," *Journal of Economic Literature*, September 1979, 17, 923-69.
- _____, and Raymond, Jennie, "A Profile of Senior Economics Majors in the United States," *American Economic Review Pro-*

- ceedings*, May 1984, 74, 19-25.
- Watts, Michael and Lynch, Gerald, "The Principles Courses Revisited," *American Economic Review Proceedings*, May 1989, 78, 263-68.
- Williams, David et al., "University Class Size: Is Small Better?," *Research in Higher Education*, No. 3, 1985, 23, 307-17.
- Yates, Judith, "Research in Economic Education: Are Our Horizons Too Narrow?," *Journal of Economic Education*, Fall 1978, 10, 12-17.

The Third Edition of the Test of Understanding in College Economics

By PHILLIP SAUNDERS*

The *Third Edition of the Test of Understanding in College Economics* (TUCE III) consists of two 33-item, four-option multiple choice tests. One test covers macroeconomics, and the other covers microeconomics. The last three questions on each test deal with international economics concepts. This enables instructors who do not cover international economics in the macro course (or in the micro course) to easily omit these questions, and use the first 30 questions as a "pure" macro (or micro) test. Each test is designed to be administered in a class period of 45–50 minutes, and copies are available in packages of 25 each from the Joint Council on Economic Education (JCEE), 432 Park Avenue South, New York, NY 10016. The tests are also accompanied by an *Interpretative Manual* (written by myself, 1991) that contains detailed norming data and additional explanatory information that cannot be included in this brief report.

Like its predecessors (see Rendigs Fels, 1967; Arthur Welsh and Fels, 1969, and myself, Fels, and Welsh, 1981), *TUCE III* has two primary objectives: 1) to serve as a measuring instrument for controlled experiments in the teaching of introductory economics at the college level; and 2) to enable instructors of particular introductory courses to compare the performance of their students with that of students in other colleges and universities. Also like its predecessors, it is a joint effort of the American Economic Association's standing Committee on Economic Education and the JCEE, which raised the funds for the *TUCE* revision.

The committee responsible for selecting, writing, and editing the questions used in

TUCE III consisted of myself as Chair; Rendigs Fels, Vanderbilt University; William Walstad, University of Nebraska-Lincoln; Michael Watts, Purdue University; and Arthur Welsh, Penn State University. We began working in the spring of 1988, and developed a pool of 85 questions to be field tested during the fall and spring terms of the 1988–89 academic year. The field test questions and some preliminary field test data were reviewed by a panel of economists in December 1988.¹ The comments of the review panel and other economists and the complete field test results were then used to prepare the 33-item tests that were normed during the 1989–90 academic year.

The content and cognitive specifications for *TUCE III* (explained in detail in the *Interpretative Manual*) are very similar to the ones used for *TUCE II*, except for the previously mentioned addition of 3 questions dealing with international economic concepts at the end of each test. Despite the similarity in overall specifications, however, only 3 of the questions on the macro test and 4 of the questions in the micro test have exactly the same wording in *TUCE III* as in *TUCE II*. Nine of the macro questions and 12 of the micro questions are completely new, and the other *TUCE III* questions are revised or edited versions of questions that were on *TUCE II*. In the remainder of this paper I will briefly elaborate on *TUCE III* tests specification, item construction, the norming sample, and a preliminary analysis of the norming data. I should also point out that, in addition to the matched

*Professor of Economics, Indiana University, Bloomington, IN 47405. A longer version of this paper will appear in the summer 1991 issue of *The Journal of Economic Education*.

¹Members of the review panel were William J. Baumol, New York and Princeton universities; Edward G. Boehne, Federal Reserve Bank of Philadelphia; Marianne Ferber, University of Illinois; Kenneth H. Miltzer, AT&T; Leonard Silk, *New York Times*; and Gary Stern, Federal Reserve Bank of Minneapolis.

norming data discussed below, we have additional *TUCE* III responses from students who took only a pretest or only a posttest, and most participating schools have sent us a set of instructor questionnaire responses, a set of student questionnaire responses, and information about individual students' SAT or ACT scores and course grades that is now being coded and matched with *TUCE* III performance data. When completed, we will have a large data set that will contain a wealth of information that should permit much more detailed analyses of *TUCE* III performance than is possible at this time. Computer disks containing all data collected will be made available to researchers by the JCEE as soon as they are compiled.

I. Specifications

Each *TUCE* III question is simultaneously classified into one of six broad macro or six broad micro content categories and one of three broad cognitive categories. The main purpose of these categories is to insure coverage of the content of "typical" principles of economics courses, and to continue the *TUCE*'s traditional emphasis on testing the *application* of basic concepts and principles. Given the scope of most principles courses and the limited number of questions that can be thoughtfully answered in a single class period, several questions on *TUCE* III deal with more than one concept or principle. This makes simple content classifications difficult, and some users may want to develop classification schemes different from the ones used to develop the overall specifications for *TUCE* III.

Since individual questions in different content categories vary in cognitive classification and difficulty, no attempt should be made to make sweeping generalizations about student knowledge of individual concepts or principles based on a single question or even a small number of questions on *TUCE* III. These tests are designed so that the *total raw score* can serve as a *general* measure of economic understanding. If individual researchers or individual instructors find that the fixed-weight content specifications of these tests are not appropriate

for their circumstances, we hope that the detailed norming data provided in the *Interpretative Manual* will enable them to interpret results in different situations in a meaningful way.

The three broad cognitive categories used to classify questions on *TUCE* III are: Recognition and Understanding (RU); Explicit Application (EA); and Implicit Application (IA). One-third of the questions on each test are in each category, and roughly half of the application questions (10 macro questions and 11 micro questions) are classified as "realistic." A realistic question is defined as one that uses a quotation taken or adapted from an actual published source or a "manufactured" quotation that might easily have appeared in such a source. In 3 questions, an actual situation is described in the stem without quotation marks being used.

The main purpose of the overall cognitive specifications is to insure a large number of questions that require students to go beyond memorization and recall, and to *use* and *apply* economic concepts. Whether or not the mental processes used by students to answer these questions actually correspond to their cognitive classification cannot be known with certainty; and any question for which a student has seen the correct answer can become a memory or a recall question, regardless of its cognitive classification. Nevertheless, we think that under normal circumstances the *total raw score* on *TUCE* III can be a useful measure of a student's *general* ability to understand and apply economic terms, concepts, and principles. As a final point on the cognitive categories used for *TUCE* III, it should be emphasized that there is no necessary or direct relation between the type of thinking or type of knowledge being tested and the difficulty of a particular question. There are hard questions as well as easy questions in each cognitive category.

II. Item Construction

Two sample questions are shown below to illustrate how individual test items are constructed on *TUCE* III. The data following

each sample question approximate overall mean performance and show the percentage of students in our norming samples selecting each alternative before (pre) and after (post) they took a principles of economics course; the point biserial correlation between the mean score of those selecting the correct alternative (shown in boldface) on that question and the mean score of the total norm group on the appropriate form of *TUCE* III (R_t); and the point biserial correlation between selecting the correct alternative on that question and the mean score on the other questions in the same cognitive category on the appropriate form of *TUCE* III (R_c).

Macro Question #8. Cognitive category "RU."

In comparing an increase in government spending on goods and services to an increase in private investment spending, we can correctly say that in the short run:

- A. they will both shift aggregate supply.
- B. they will both shift aggregate demand.**
- C. government spending is inflationary; private investment is not.
- D. government spending must equal taxes; private investment must equal saving.

Pre	Post
25%	17%
29%	50%
22%	15%
24%	18%
$R_t = .20$	$R_t = .38$
$R_c = .33$	$R_c = .44$

Micro Question #9. Cognitive Category "EA."

If all of the firms producing a product in a perfectly competitive market are required to adopt antipollution devices that increase their cost of production, one would expect:

- A. the demand for the product to fall.
- B. the market supply curve to shift to the left.**

- C. the long-run economic profits of the individual firms to fall.
- D. the short-run economic profits of the individual firms to remain unchanged.

Pre	Post
11%	9%
39%	56%
36%	25%
14%	10%
$R_t = .33$	$R_t = .44$
$R_c = .43$	$R_c = .50$

Both of these questions illustrate the point that, unless there was a strong reason for doing otherwise, all of the alternatives on each question of *TUCE* III are arranged uniformly from the shortest to the longest. This was done so that the longest alternative, which some "test-wise" students may think is usually the correct alternative, does not "stick out" and call attention to itself. Further, special care was taken to insure that each of the alternatives (A, B, C, or D) is the correct option the same number of times (8) with the exception that, on a 33-item test, one option has to be correct 9 times. On the macro *TUCE*, the "extra" option is "A," and on the micro *TUCE*, it is "D". This again was done on purpose, since another rule of thumb that test-wise students sometimes use is that examiners often like to "hide" or "bury" the correct alternative in the middle of the sequence rather than expose it in one of the first or last positions. To the maximum extent possible, we sought to make the overall *TUCE* score a measure of students' economic understanding rather than multiple-choice test-taking skills.

Both of the sample questions have good statistical properties: 1) students did better on the posttest than on the pretest; 2) all the alternatives were plausible and attracted some student response; 3) the point biserial correlation (R_t) between the mean score of students choosing the correct answer and the mean total test score is relatively high; and 4) the point biserial correlation with other items in the same cognitive category (R_c) is higher than the correlation

with the total test score (R_i). Although it is not shown, both questions also had the desirable statistical property that each incorrect answer was negatively correlated with the total test score, that is, the point biserial correlation between the mean score of those selecting each incorrect answer on each question and the mean score of the norm group for the total test was negative. Also, the percentage of students choosing each individual incorrect alternative in these questions is lower on the posttest than on the pretest. A careful analysis of the detailed item data in the *Interpretative Manual* reveals that most of the individual items on *TUCE* III possess all six of these desirable statistical properties.

III. Norming Sample

The sample of schools used to collect norming data for *TUCE* III is purposely larger and broader than the sample used to norm previous editions. Two-year colleges, at which large numbers of students take principles courses, are included for the first time, and we have broader representation of other types of schools as well. Fifty-three schools provided us with enough data to compile a matched sample of 2,724 students who took the macro test at both the beginning (pretest) and the end (posttest) of the course, and a similar matched sample of 2,726 students for the micro *TUCE*. Forty-one schools provided us with both macro and micro data. Four schools provided us with macro data only, and 8 schools provided us with micro data only. The 2,724 matched macro student responses came from 92 different sections at 45 schools. The 2,726 matched micro student responses came from 96 different sections at 49 schools.

Since some sections and some schools in the norming sample used only the first 30 questions on the posttests and others used all 33 questions, our norming sample for the last 3 international questions on each test is smaller than on the first 30 questions. For the 33-question macro test, the sample is 1,324 students from 55 sections at 33 schools.

For the 33-question micro test, it is 1,426 students from 52 sections at 28 schools. Since there are some differences in the sections and the schools using 33 questions and those using only 30 questions, extra care must be used in comparing student performance on the last 3 questions with performance on the first 30 questions on both macro *TUCE* III and micro *TUCE* III. To help avoid confusion on this issue, separate sets of detailed tables have been prepared for the 30-question samples and the smaller 33-question subsamples. It is also worth repeating that data being compiled from the instructor questionnaires and the student questionnaires will later permit much more sophisticated analyses of factors affecting *TUCE* III performance than is possible at this time.

While great pains were taken to obtain a *representative* sample for norming *TUCE* III, neither the macro sample nor the micro sample is a statistically *random* sample. Some schools that were initially asked to participate in the norming chose not to do so, and some schools that agreed to participate failed to provide a complete set of data for a variety of reasons. Nevertheless, compared to the 25 schools that participated in the norming of *TUCE* I and the 36 schools that participated in the norming of *TUCE* II, the 53 schools in the *TUCE* III sample are considerably more representative of the wide range of higher education institutions in the United States today. Using broad Carnegie Council categories, 12 of the norming schools are doctorate granting institutions, 23 are comprehensive universities and colleges, 13 are four-year and liberal arts colleges, and 5 are two-year colleges.

IV. Preliminary Analysis of Norming Data

Table 1 shows the pre- and posttest mean raw scores and standard deviations on the 30- and the 33-question versions of macro *TUCE* III and micro *TUCE* III. The posttest mean scores of 14.31 out of 30 (47.7 percent) and 15.15 out of 33 (45.9 percent) on macro *TUCE* III and 15.36 out of 30 (51.2 percent) and 16.67 out of 33 (50.5 percent)

TABLE 1—PRE- AND POSTTEST PERFORMANCE
ON *TUCE* III

Test Form	N	Mean Raw Score ^a	
		Pre	Post
Macro			
30 Questions	2,724	9.18 (3.05)	14.31 (5.24)
33 Questions	1,324	10.57 (3.45)	15.15 (5.40)
Micro			
30 Questions	2,726	10.71 (3.88)	15.36 (5.67)
33 Questions	1,426	12.35 (4.59)	16.67 (6.25)

^aStandard deviations are shown in parentheses.

on micro *TUCE* III are lower than the comparable mean scores on the previous editions of the *TUCE*. Some of this difference may be due to a difference in the questions used, and some may be due to the inclusion of a broader range of schools and students in the norming sample. It does not necessarily imply that we are now doing a worse job of teaching introductory economics than we used to, but that may be a hypothesis that is worth investigating.

The pre-to-post gains of 9.18 to 14.31 (55.9 percent) on the 30-question macro, 10.57 to 15.15 (43.3 percent) on the 33-question macro, 10.71 to 15.36 (43.4 percent) on the 30-question micro, and 12.35 to 16.67 (34.9 percent) on the 33-question micro tests indicate that, on average, students do improve their understanding of college economics during their principles courses. We hope that the instructor and student information gathered during the norming process will help us identify the factors that contribute to this improvement in a way that can help us increase it.

One test statistic concerning *TUCE* III that is not shown in Table 1, but which is worth mentioning, concerns test reliability. Test reliability refers to consistency of measurement. Any test score contains a certain degree of error of measurement. For this reason, if a student were tested a number of times, the scores would be expected to vary.

A reliability coefficient may be thought of as an estimate of the correlation between scores obtained on one test at one time and those obtained on a similar test at a different time. Kuder-Richardson Formula 20 (KR20) has been used to estimate the reliability of all editions of the *TUCE*. The posttest KR20's for micro *TUCE* III (.80 for 30 items and .81 for 33 items) are higher than the reliability coefficients for any of the previous micro forms. The posttest KR20's for macro *TUCE* III (.77 for 30 items and .75 for 33 items) are comparable to the reliability coefficients of most of the previous macro forms, but not quite as high as *TUCE* II Macro Form A, which had a coefficient of .81.

As the reliability data indicate, and as most economists would expect, we found it more difficult to construct macro *TUCE* III than micro *TUCE* III. There is less consensus in the profession on macro principles than on micro principles, and this poses real problems for test construction and interpretation. Individual instructors who want to use *TUCE* III data to analyze the performance of their students should first examine the tests and their detailed specifications carefully to make sure that they are congruent with the content and purposes of their own courses. In cases where some questions do not seem appropriate, the detailed item data provided in the *Interpretative Manual* can be used to compare the percentage of students answering individual questions correctly. Given the wide range of students and schools contained in the norming sample, the detailed data contained in the *Manual* may also be more useful than a simple comparison of raw scores or percentages.

If it is determined that the overall specifications on *TUCE* III are appropriate for a particular situation, for example, the percentile data (not shown here) can be used to put scores in context in the following way. A class with a macro pretest mean score of 11 and a posttest mean score of 18 on the first 30 questions would be in the 75th percentile of individual students in both cases, whereas a class with a pretest mean score of 9 and posttest mean score of 16 on the first 30

macro questions would be in the 52nd percentile on the pretest and the 65th percentile on the posttest. Thus, a pre-to-post gain in mean scores of 7 questions would be interpreted differently in these two situations. Likewise, a class with a pretest mean of 11 and a posttest mean of 16 on the 33-item micro *TUCE* would be in the 43rd percentile of individual students on the pretest and the 49th percentile on the posttest. This would indicate that the average performance of students in this class had increased relative to the national norms on *TUCE* III. Alternatively, a pretest mean of 11 on the 33-item macro *TUCE* (60th percentile) and a posttest mean of 16 (60th percentile) would indicate that the average performance of the students had stayed constant relative to the national norms.

REFERENCES

- Fels, Rendigs, "A New Test of Understanding in College Economics," *American Economic Review Proceedings*, May 1967, 57, 660-66.
- Saunders, Phillip, *Third Edition of the Test of Understanding College Economics: Interpretative Manual*, New York: Joint Council on Economic Education, 1991.
- _____, Fels, Rendigs, and Welsh, Arthur L., "The Revised Test of Understanding College Economics," *American Economic Review Proceedings*, May 1981, 71, 190-94.
- Welsh, Arthur L., and Fels, Rendigs, "Performance on the New Test of Understanding in College Economics," *American Economic Review Proceedings*, May 1969, 59, 224-29.

Providing Earth Observation Data from Space: Economics and Institutions

By MOLLY K. MACAULEY AND MICHAEL A. TOMAN*

Burgeoning interest in understanding global environmental change has greatly increased demand for remote sensing of the earth from space.¹ Over 60 percent of a \$1 billion U.S. budget to study global change during fiscal year 1991 was allocated to space-based environmental data collection, including down payment on a \$52 billion space data program, the "Earth Observing System" (EOS), planned for the next 15 years. This substantial expenditure, together with the huge technological scale of EOS (it will consist of large orbital platforms each carrying a multitude of sensors), have promoted considerable debate and alternative proposals for smaller-scale, modular spacecraft each carrying one or a few sensors. Proponents of the alternatives hold that the smaller-scale approach, in addition to being less expensive, foregoes few scale economies and avoids the "eggs-in-one-basket" risk of EOS, in which failure in launching a spacecraft or failure of one instrument on a

multi-instrument spacecraft causes failure of the whole system. Advocates of EOS are concerned that the small-scale approach not only fails to exploit scale economies but may yield much lower information content, claiming that significant economies of scope are obtainable in operating and taking measurements from sensors aggregated on large spacecraft.

In this paper we first ask whether scale and scope economies appear substantial enough, and risk small enough, to justify the large-scale approach presently envisioned.² We then ask two separate but related questions about the organization of the project, namely whether these supply attributes require the technology to be an almost exclusively governmental activity. A dearth of information about remote sensing economics forces our approach to the production technology to be heuristic rather than empirical. Subject to this constraint, we fail to find evidence of significant scale and scope effects in large spacecraft and sensor manufacturing and launch, even after taking account of risk differentials among larger-scale and smaller-scale systems (the former

[†]*Discussants:* Arthur DeVany, University of California-Irvine; Nancy Rose, MIT; James Dearden, Lehigh University.

*Fellow and Senior Fellow, respectively, Resources for the Future, 1616 P Street NW, Washington, D.C. 20036. We thank Doug Bohi, Roger Noll, Tom Hazlett, and James Dearden for comments.

¹The vantage point of space offers synoptic and simultaneous measurement of atmospheric, land, and oceanic processes. Space remote sensing measures energy reflected by the earth's surface or atmosphere; water, vegetation, clouds, and other elements have different reflective "signatures" and their measurement adds detail to understanding environmental processes (i.e., temperature change, ozone depletion). We note that data from aircraft, balloons, and analysis of ice, sediment, and other ground processes are sources of "intermodal" competition (and complementarity, as some "ground-truthing" is needed for interpreting space data).

²We acknowledge the political economy arguments that predict overwhelming political momentum for lumpy, large-scale projects even if they are economically cost ineffective (see, for example, Linda Cohen and Roger Noll, 1991, and Macauley, 1990). But the highly visible record of the nation's space program in unsuccessfully operating large projects (difficulties with the shuttle, the Hubble Space Telescope, and the space station) is already resulting in considerable rethinking of the scale of EOS (for example, see National Research Council, 1990; Ferris Webster, 1990; Bob Davis, 1990; William Ganoe, 1990; Richard Stevenson, 1990; and William Hansen et al., 1990). Our analysis is by no means definitive, but it may illustrate key considerations for the political debate.

involves low probability, high-cost consequences; the latter involves somewhat higher probability, low-cost consequences). Thus we conclude that the case appears strong for consideration of smaller-scale systems. In addition, our study suggests several implications for the role of the private sector, even if the remote sensing data are considered a public good. These suggestions range from easing entry barriers to participation in the supply of the technology and reforming data access and pricing policy to issuing "information vouchers" to scientists involved in research on environmental change.

An important caveat is that our research does not assess the value of remote sensing information, where that value is reflected in the consequences of actions taken or not taken in response to the information. The value could be in the billions of dollars (see Alan Manne and Richard Richels, 1990, and William Nordhaus, 1990, for general discussions of information value regarding climate change). Savings in social costs attributable specifically to space-based information, however, would only be a part of the total value of information, and would depend on the space data being received and interpreted in a timely enough fashion to affect decision making (Marcia Smith and John Justus, 1990, emphasize this point). A better understanding of the value of the heterogeneous output provided by remote sensing is an obvious extension of the research in this paper.

I. The Supply of Space-Based Remote Sensing

A stylized representation of the supply of earth observation and information processing includes four vertically related stages: 1) spacecraft and sensor manufacture; 2) spacecraft system launch; 3) sensor system operation, signal transfer to ground stations, and primary processing of information; and 4) "value-added" processing and dissemination of imagery. Our 1990 paper contains details of our analysis of scale and scope effects in these stages; a summary follows.

There is significant reason to believe that stage 4 approximates a constant returns

technology given the modularity of workstations and software. Stages 1 and 2 exhibit some scale effects from the spreading out of fixed R&D and facilities costs (in stage 1) and launching larger-mass payloads (stage 2). Several factors attenuate these economies, however. Most notably, economies from multiple production are likely to be limited. Although the 15-year time frame envisioned for space-based environmental monitoring by EOS will require replacing spacecraft during the period, replacement is to occur only roughly every 3 to 5 years. Thus even if scale economies are present, they may not be exploited in small-scale production runs.

The principal issue in system launch is the tradeoff between launching a large platform with many sensors or dividing the sensor suite among several smaller platforms. There are some scale economies in launching larger-mass payloads (see our 1989 article). However, dividing the sensor suite among several launches may reduce the risk of system failure from an unsuccessful launch and also provide flexibility in subsequently replacing or upgrading components of the sensing system.

Potential economies in stage 3 hinge partly on whether larger systems can be operated at lower cost or provide higher quality data than smaller systems. Conventional wisdom holds that simultaneous measurement by different sensors is required for mutually referenced and coordinated data. However, alternative methods distribute sensors over a large number of smaller spacecraft, steer these in tandem, and filter out referencing errors in more intensive ground processing. Other economies at this stage center on transmission and initial processing of the data. These operations may be more economical on large rather than small spacecraft if scale and scope effects are obtainable by consolidating total power requirements.

How large the potential economies in stage 3 must be to justify a large-scale approach such as EOS is suggested by available information about the cost of various alternatives. EOS requires 6 large spacecraft and some 90 instruments; proposed

alternatives involve as many as 60 smaller spacecraft and 60 to 75 instruments. A comparison of the ratio of hardware and launch costs to the number of instruments (a rough measure of a system's information content) suggests significant diseconomies of scale with systems involving large spacecraft or a large number of medium spacecraft. While absolute costs may be understated (as is typical of space projects), we suspect that relative costs of alternatives are good approximations to relative orders of magnitude. On this relative basis, the larger systems are about three times more expensive than more modest alternatives. If this observation is correct, then the more costly approaches must offer some combination of significant operating economies and higher-quality data to justify these substantial hardware and launch cost differences.

Using available engineering data, rough calculations of the cost magnitudes involved in launch or instrument failure reduce the cost effectiveness of some of the smaller-scale alternatives (essentially because these systems require the success of more events: in one case, 60 launches). However, our probability calculations ignore differences in the relative importance of various instruments. For example, the loss of data could be minor if less critical instruments failed. An important extension of our analysis would involve calculating more accurate estimates of risk premia that incorporate weights based on the expected information content of various components of the proposed systems.

II. Policy Issues

Despite the lack of conclusive empirical information, enough hints about the workability of smaller-scale, more modular systems can be found in the technical literature to suggest that this option warrants serious consideration before the substantial sunk costs of a system the size of EOS are incurred. If small-scale systems are viable, then much of the argument for an exclusive role of government to remedy market failure due to scale and scope effects in the supply process is undermined. (Even if the

data provided by the program are viewed as pure public goods, then the private sector might be beneficially involved to a greater extent than currently planned.) At least three policy implications follow from these observations.

First, entry by smaller-scale operators serving niche markets (for example, hydrologic forecasting, forest inventory) could be possible without undue duplication of facilities or problems of coordination in the use of fixed facilities (say, launch sites). However, encouraging entrepreneurial entry would require the dissolution of present entry barriers, including legal restrictions on remote sensing data access and discriminatory pricing. In addition, ready access would need to be provided (at user cost) to existing public facilities like launch sites and communications networks.

Second, even if scale economies are deemed important enough to require a large-scale system, the U.S. government could encourage entrepreneurial entry by operating the system as a joint venture (again, after reducing artificial entry barriers). Any operator could locate a sensor package in the system by paying the appropriate cost that the entry would cause the system to bear.³ This approach combines competitive rivalry in serving end-use markets with the reaping of any economies from larger-scale facilities. (Some coordination among independent operators would still be required to achieve efficiency in the use of fixed facilities; for incentive compatible solutions to coordination in the case of the space station, see Jeffrey Banks et al., 1989.)

A third possibility is to provide subsidies in the form of "information vouchers" to those seeking earth observation information for stipulated public good purposes. Recipients could use the vouchers to procure information they deem most suited to their needs (including non-space-based data) and would share in any resulting cost savings

³Some evidence of private sector willingness to sponsor sensors has been documented in a recent NASA program, the Earth Observation Commercialization Applications Program.

(Macauley, 1989, applies a voucher approach to the supply of launch services to the science community). This scheme might be difficult to use if scope economies turn out to be so substantial that voucher holders could succeed only by seeking their own joint ventures with other users or suppliers. Yet the resulting transaction costs might be manageable (and they would most likely not exceed transactions costs already implicit in the centralized approach to EOS now being taken) or the scope economies might not be so substantial. If feasible, this approach could provide a useful dose of economic incentives to users and suppliers of earth observation information.

REFERENCES

- Banks, Jeffrey S., Ledyard, John O. and Porter, David P., "Allocating Uncertain and Unresponsive Resources: An Experimental Approach," *RAND Journal of Economics*, Spring 1989, 20, 1-25.
- Cohen, Linda and Noll, Roger, *The Technology Pork Barrel*, Washington: Brookings Institution, 1991.
- Davis, Bob, "Report Provides Support for the Critics of Using Big Satellites to Study Climate," *The Wall Street Journal*, August 22, 1990, B4.
- Ganoe, William H., "Earth Probes," *Ad Astra*, May 1990, 47.
- Hansen, James, Rossow, William and Fung, Inez, "The Missing Data on Global Climate Change," *Issues in Science and Technology*, Fall 1990, 62-69.
- Macauley, Molly K., "Launch Vouchers for Space Science Research," *Space Policy*, November 1989, 5, 311-20.
- _____, "The NASA Budget: For What, For Whom, and How Big?," paper presented at the 3rd Colorado Space Policy Conference, Boulder, August 1990.
- _____, and Toman, Michael A., "Eye in the Sky: The Economics of Space Remote Fencing," RFF Discussion Paper ENR 91-06, Washington, 1990.
- Manne, Alan S. and Richels, Richard G., "Buying Greenhouse Insurance," mimeo., November 1990.
- Nordhaus, William O., "To Slow or Not to Slow: The Economics of the Greenhouse Effect," paper presented at the annual meeting of the American Association for the Advancement of Science, 1990.
- Smith, Marcia S. and Justus, John R., *Mission to Planet Earth and the U.S. Global Change Research Program*, CRS Report for Congress, 90-300 SPR, Washington: The Library of Congress, CRS, 19 June 1990.
- Stevenson, Richard W., "Scientists Back Smaller, Simpler Satellites for Monitoring Climate," *The New York Times*, August 22, 1990, B4.
- Toman, Michael A. and Macauley, Molly K., "No Free Launch: Efficient Space Transportation Pricing," *Land Economics*, May 1989, 65, 91-99.
- Webster, Ferris, "EOS, the Earth Observing System: NASA's Global Change Research Mission," paper presented at the 3rd Colorado Space Policy Conference, Boulder, August 1990.
- National Research Council, *The U.S. Global Change Research Program*, Washington: National Academy Press, 1990.

Trading Orbit Spectrum Assignments in the Space Satellite Industry

By HARVEY J. LEVIN*

Is it feasible to trade rights to use orbital slots and associated satellite frequencies in a market? Can such rights be delimited today? What is the role of international organs like the International Telecommunications Union (ITU), national agencies like the Federal Communications Commission (FCC), or the United Kingdom's Frequency Planning Organs (FPOs)? What evidence is there that such rights or assignments are in fact configured now and do indeed exist? Are they frequently traded in international as well as domestic broadcast and mobile radio services?

Even if practical and viable, what economic policy or equity purposes are served by configuring actual markets for transferable or exchangeable orbit spectrum assignments? Or for schemes where national or international authorities distribute such assignments under competition or specifically by auction to the highest bidder? That is, through public as well as private auctions? Do auctions in any case further distributive equity and not just economic efficiency?¹

I. The Tongasat Episode

The latest evidence on spectrum value probably derives from the Kingdom of Tonga's proposed orbit slot auctions. In fact, all 16 "slots" Tonga laid claim to this past year are still subject to approval by ITU's International Frequency Registration Board

(IFRB). (See Edward Andrews, 1990, p. A-1.) Today, the estimated asking price is roughly \$2 million per slot, per year (see Andrews, p. C-17). Even a few years back, comparable references could have been made to a growing number of then-proposed joint space ventures, and later of slot auctions.

A still unanswered question is whether Tonga literally sought applications for all 16 remaining slots over the Pacific to transform the ITU registration process into a mechanism for financial speculation in the geostationary orbit. And, whether, as a consequence, those applications will operate to subvert the stated purpose of ITU's Radio Regulations, or of the Union itself, or, in particular, of Articles 4 and 29 of the ITU Constitution.

Of special interest were a number of "unsolicited offers to Intelsat, and undoubtedly to others, proposing lease or other arrangements for...use of the slots." As perceived by Intelsat, Tongasat's press releases, offering statements, and mass filings with IFRB are mainly geared to "gain...control over an excessive number of orbital slots which could then be speculated in or sold for financial gain." (Letter from Intelsat Director General Dean Burch to Members of IFRB, June 12, 1990, p. 1.) The main issue is whether Tongasat is applying for orbital slots far in excess of its projected economic needs. If so, is the orbital arc being "claimstaked" to prevent "other ITU members from registering and using the orbital slots for legitimate purposes without financial remuneration to Tongasat"? (Letter, p. 2.) Does Tongasat intend to distribute "orbital slots by selling or auctioning them to the highest bidder"? If so, is this "hoarding of orbital locations and associated frequency spectrum...contrary to the spirit of Article 29 of the ITU Constitution (Nice, 1989)"? (Letter, p. 2.)

*University Research Professor and former Augustus B. Weller Professor of Economics, Hofstra University, Hempstead, NY 11550.

¹For recent evidence on spectrum value and market-type transactions in New Zealand, see *Treasury Submission...*, 1985; Robin Foster et al., 1988; Charles Jackson and Foster, 1989; Nancy Kuehl, 1990. On the economics of spectrum allocation in other countries outside the United States, see OECD, 1990; CSP International, 1987; *Wall Street Journal* editorial, 1990; and Government of Canada, 1990.

There is in any case a widespread international consensus apparent in the Radio Regulations and many ITU resolutions that the geostationary satellite orbit, and associated radio frequencies, are without question scarce natural resources owned by no one nation but rather by all. These resources are subject to standards of economic efficiency and equitable allocation among all nations in meeting their legitimate communication needs. The big question is whether the slot and spectrum auctions envisioned by Tongasat will facilitate or impede the furtherance of equity and efficiency as criteria for satellite regulation and orbit spectrum management.

II. The Pacstar Joint Venture

Let us take note of the recent proposed joint venture between the TRT Telecommunications Corporation's subsidiary Pacific Satellite Inc. (PSI) and the Papua New Guinea (PNG) Public Telephone and Telegraph Co.² There, PNG notified the IFRB 2 years beforehand, *on behalf of PSI*, that PSI would use two PNG orbital assignments in return, giving PNG free domestic circuits in return. In this scenario, the developed country (DC) entity virtually *borrowed* an unused developing country (LDC) assignment to illuminate part of the DC's market. In return, it would pay the LDC in kind with *free* transponders or reduced-rate domestic circuits, at low incremental cost to the DC firm, once the latter's satellite was in place.

In all such cases one could say that LDCs are already finding ways to lease out, lend, or otherwise trade their orbit spectrum assignments; and to do so for a portion of Intelsat's revenues (14 percent return on investment); or for a comparable return on regional investment by an Arabsat, Andeansat, or Euthelsat.

The advantage to Pacstar in using PNG's assignment is twofold. First, it would pre-

sumably get "better" (less costly) coordination under ITU with joint support and sympathy in the South Pacific. Second, it might thus avoid being thrown into the large pool of applicants at FCC, including proposals for a North Atlantic service. Processing that full queue would take a long time, and lots of money. If Pacstar had to wait its turn on line it might well have far more trouble in getting preferred assignments than if PNG applied for them directly instead.

The engineering expertise, and *intimate knowledge of international regulatory-coordination procedures* that Pacstar brought into the joint venture, plus avoidance of the FCC queue, have doubtless resulted in important cost savings. These were in turn exchanged for free domestic satellite circuits for PNG. Hence PNG's slot value may have been created by international administrative sluggishness or simple incompetence in dealing with ITU and IFRB direct.

In the Pacstar case, PSI wanted to create a system in the South Pacific and arranged for its origin spectrum assignment in a then-unique way. Instead of waiting in line at FCC, the ITU, or IFRB, PSI sought to avoid such delays by entering directly into a Joint Venture with a third party, PNG, or more accurately its PTT.³ As part of the bargain, PSI would give PNG one free transponder in perpetuity, and maybe the Tracking, Telemetry & Control Module (TTCM) too, though not the cost of operating it. So the question is what it would cost PNG to *operate* the TTCM for its 7-year lifetime, compared to the cost of *buying* or *renting* one transponder for the same period.

III. The Panamsat-Peru Episode

In this case, the Panamsat Corporation wanted to provide diverse programming and service in the Western Hemisphere. Under FCC and ITU regulations, however, it could not gain access to the orbit spectrum re-

²On new interests and insights in satellite joint ventures and consortia and related arrangements, see The Transmart Company, 1990; UN Center on Transnational Corporations, 1988.

³Illustrative joint ventures between entities in LDCs and the United States by now include Brazil, Venezuela, Dominican Republic, Guyana, Haiti, India, and the Malaysian Republic. (See Carl Christol, 1975.)

quired. Without a foreign correspondent, that is, someone to communicate with at the other end, the corporation would simply have a right it could not use.

As traced in the trade press, Panamsat did locate a correspondent when Peru came forth. Again, as with Pacstar, Peru was paid one free transponder per year, in perpetuity, a gift worth \$1 or \$2 million annually had she had to rent it outright. Because Peru needed only one-fifth of a transponder, however, she got a little extra, and here, too, was an economic benefit of sorts.

IV. The Panamsat-Cygnus Transaction

Cygnus and Panamsat were both satellite companies. However, Cygnus was poised, anxious and authorized to serve over the Atlantic to Western Europe and the mid-East, though with no satellite equipment in orbit. On the other hand, Panamsat's hybrid satellite had 24 low-cost C-band transponders which she was already authorized to use. She also had another 12 transponders higher up in Ku-band, where it was more costly to operate, and where FCC continued to prohibit use.⁴

Accordingly, one could say that Panamsat had the needed *equipment*, while only Cygnus had the *rights* to operate in Ku-band at the same orbital location. So that Panamsat had the physical capability and the equipment already in place, while Cygnus had the rights to use all her 24 planned transponders, but lacked the equipment needed to start. Because starting from scratch at that time, and then building the needed satellite would simply take too long, Cygnus would never have made the deadline. Therefore, the two companies might have entered into some sort of agreement with one company buying the other out, or both companies creating a joint venture.⁵

⁴The FCC did not want Panamsat to use only half of its Ku-band satellite (12 transponders), preferring instead that she use a full C-band satellite of 24 transponders.

⁵See FCC, 1987; see, especially, paras. 6, 9-12.

Cygnus did in any case have advertising contacts, marketing services, client lists, and goodwill. The buyer (Panamsat) presumably had to pay for all this, and, of course, she also wanted to remove a potential rival from the field. In addition, there was something of potential value left in the bankrupt Cygnus, and the challenge was to resuscitate it when Cygnus herself could not do so.

Nevertheless, if all that remained after bankruptcy was Cygnus's bare authorization, the FCC would presumably approve the transfer at a price recovering the seller's cost of securing the original spectrum assignment (see FCC, para. 6). The transfer price would be roughly \$300,000-\$350,000. This would cover Cygnus's reasonable out-of-pocket costs including goodwill and marketing information, but nothing more, to avoid any proscribed trafficking in licenses.

V. Atlantic Star

Here the partners would be Hughes and the state of Ireland with the proposed joint venture geared to operate a direct broadcast satellite. Ireland's orbit spectrum assignment would presumably be used, thereby saving Hughes the time and money she must expend should she proceed through FCC to the ITU (see Hughes Communications, Inc., 1990). So the proposed Atlantic Star seems similar to Pacstar. In both cases, a large U.S. company (TRT or Hughes) contributes capital and know-how to a joint venture (Pacstar, Atlantic Star), whereas a remote developing island nation (Papua New Guinea), or a small European economy (Ireland), contributes the needed orbit spectrum assignments.

With Atlantic Star, Hughes proposed to hold 80 percent of the investment, and Ireland, 20 percent. The main use was to be for English or Irish language programs for Europe, the Irish getting in because of dissatisfaction with British plans and activity. But now that the British have approved their own system, the Irish seem to have faded out. So if Hughes can get contracts from both the United Kingdom and Ireland that would suffice, and there'd be no need for Atlantic Star. (See Peter Liska, 1987.)

IV. Conclusion

In addition to the five well-known episodes just discussed, market-type transactions, spectrum value, and joint ventures also figure in other cases. These involve additional joint ventures such as Japan-Communications Satellite, Asiasat, Rupert Murdoch's Sky TV-Disney Channel, or his joint venture with Luxembourg, where the latter was to provide the orbital assignment and Murdoch the broadcast satellite. Or where RCA or GT&E leases or rents unused Anik-C capacity paying Canada \$2 million U.S. annually to tilt the satellite southward until Canada herself might need that capacity in her northern regions. For several prior years, the Anik capacity enabled those companies to perform reliably before their own new equipment was orbited and operating.

REFERENCES

- Andrews, Edmund L.**, "Tiny Tonga Seeks Satellite Empire in Space," *New York Times*, August 28, 1990, A-1, C-17.
- Christol, Carl**, "Space Joint Ventures—The United States and Developing Nations," *Akron Law Review*, Spring 1975, 8, 398–415.
- Foster, Robin et al.**, *Management of the Radio Frequency Spectrum in New Zealand*, National Economic Research Associates (NERA), London, 1988, Ministry of Commerce, New Zealand.
- Jackson, Charles, and Foster, Robin**, NERA, "The New Zealand Spectrum Project: Description and Observations," 17th Annual Telecommunications Policy Research Conference, Airlie, VA, October 1–3, 1989.
- Kuehl, Nancy**, "New Zealand Reeling in Profits from Auctions of Cellular Spectrum," *Radio Communications Report*, July 23, 1990, 9, 1, 14.
- Liska, Peter**, "Hughes Heading European Push," *Satellite Communication*, December 1987, 11, 45–46.
- CSP International**, *Deregulation of the Radio Spectrum in the U.K.*, Dept. of Trade and Industry, HMSO, 1987.
- Federal Communications Commission**, Common Carrier Docket No. 84–1299, File No. CSS-87-001-A(CP), File No. CSS-87-002-TC, Memorandum Opinion and Order, January 14, 1987, Applications for Consent to Transfer of Control (Cygnus), and for Consent to Pro Forma Assignment (Panamsat); paras 6, 9–12; Panam and Cygnus, paras 10–11.
- Government of Canada**, Dept. of Communications, Spectrum and Orbit Policy Directorate, "Towards a Spectrum Policy Framework for the 21st Century," Discussion Paper, Ottawa, September 1990.
- Hughes Communications, Inc.**, Attachments: "The Atlantic Satellites System," descriptive brochure; "A Decade of Innovation," fact sheet; Private communication from Carson E. Agnew, Vice President, Hughes Communications New Venture Organization, to H. J. Levin, Los Angeles, CA, March 5, 1990.
- New Zealand**, *Treasury Submission to the Royal Commission on Broadcasting and Related Telecommunications*, September 1985.
- OECD**, Committee for Information, Computer and Communications Policy, "The Economics of Frequency Allocation" (Note by the Secretariat), Paris, October 8, 1990.
- The Transmart Company**, *Joint Venture News—A Newsletter of International Business*, 1, Anchorage, January 1990.
- UN Center on Transnational Corporations**, *Joint Ventures as a Form of International Economic Cooperation*, New York: United Nations, 1988.
- Wall Street Journal** editorial, "Congress's Wheel of Fortune," July 1990.

Torts and Orbits: The Allocation of the Costs of Accidents Involving Spacecraft

By ANN M. BUTLER AND NEIL A. DCHERTY*

Who pays when a satellite fails to burn up in reentry and crashes in Manhattan, or when a Titan rocket fails on launch and crashes on Cocoa Beach? These are matters for courts and legislators. In 1988, the Commercial Space Launch Acts Amendments (CSLAA) capped common law liability at \$0.5 billion (with the federal government assuming the next \$1.5 billion) and made third-party insurance compulsory, as available, up to the cap. But the costs can be reallocated amongst the various parties involved in the venture. Thus "who pays" becomes an economic choice. This choice has significant implications for resource allocation and will influence the level and quality of investment in space activity. We examine economic issues surrounding the allocation of liability, particularly contract design, and the effects of insurance on market efficiency.

The main players in the commercial satellite industry are manufacturers, operators, launch contractors, and a host of supporting industries including subcontractors and insurance firms to which some risk can be transferred. To these may be added third parties who bear the risk of injury or property damage from failure of a dangerous technology, and the government whose interest stems from these negative externalities and from positive technical externalities which accrue both to private and military activities.

Liability exposure clearly shapes incentives for real investment and emphasizes the importance of contract design in allocating risk. Contract design is important because

parties differ in their abilities to bear risk and in their abilities to minimize the costs of accidents. The optimal contract trades off risking bearing against efficiency in production and loss control in the usual way.

Contract design would be simplified if insurance markets were frictionless. Contracts between operators, manufacturers, and launchers would optimally allocate liability for third-party loss, and indeed other loss, to the party who could minimize total costs. However, insurance markets are far from perfect; small numbers of satellite launches imply limited diversification for insurers and small samples from which to estimate loss distributions. Moreover, technology is constantly changing and between some losses there is a high correlation (for example, satellites on the same launch vehicle). Finally, insurance encounters significant moral hazard since the complexity of the technology makes it costly to monitor the policyholder's actions, and the small sample of events makes "reputation" an ineffective control.

A further set of issues complicates the issue of contract design but makes it more interesting. Players vary considerably in net worth, yet all face potential liabilities for third-party damage of billions of dollars. The expected gain from default on these liabilities is highest for firms with small net worth, creating an apparent diseconomy of scale. However, liability rules, such as joint and several liability, make larger firms leery of entering joint ventures with small firms. Many small firms are innovators but rely on joint ventures to supply ancillary technology, manufacturing facilities, or capital. These innovations might be lost unless small firms are able to internalize the full costs of their activities *ex ante*. This prospect defines a particular role for insurance, and against this background we can examine the effect

*Florida State University, Tallahassee, FL 32306, and Wharton School, University of Pennsylvania, Philadelphia, PA 19104, respectively. A more extensive paper is available from the authors.

of compulsory insurance and liability caps recently introduced.

I. Limited Liability, Compulsory Insurance, and Liability Caps

The common-law basis of liability for launch, reentry, and similar accidents, (for example, strict liability or negligence) is untested since there have been no such lawsuits to date. However, a clear liability exists and is internalized to the industry by common law. The size and nature of this cost depends both on liability rules and the economic characteristics of the players. The value of judgments awarded by courts rests on the ability of the plaintiffs to enforce the awards against defendants with limited net worth. With such liability costs, firm size, organizational form, and contractual allocations of liability, become strategic variables.

The commercial space market comprises players of different sizes and organizational structures, ranging from the U.S. government to individual scientists who attempt to bring to market innovative technologies based on their own research. In between are stock firms including such major players as Hughes, McDonnell Douglas, and Rockwell. The value of an award for injury to a third party depends on who the plaintiff happens to be and on the extent and quality of his insurance. First suppose that the good is produced by a single firm. We relax the assumption later.

Limited Liability with Single-Firm Production.

All producers ultimately have a limited liability for their actions. For stock firms, liability is truncated at net worth. In other organizational forms, liability may be limited by personal bankruptcy laws and the wealth of the firm's owners. We discuss stock firms, though the concepts generalize. The effect of limited liability is to create a "put" option for the firm's owners, who put the loss to the injured parties when the loss exceeds net worth (the striking price). The firm's long position in the put increases firm value. The value of this put option depends on the net worth of the firm as well as its risk characteristics. The option will be more

valuable to small and risky firms than to large and stable firms. This creates a distinct cost advantage to small firms.¹

The main implications of the cost advantage for small firms lie in organizational choices. High-risk technologies would be more likely to appear in small-size firms and existing firms would tend to "spin off" high-risk activities. In other industries, there is recent evidence that this has occurred (see A. Ringleb and S. Wiggins, 1990).

The cost advantage of small firms is unlikely to be removed by voluntary insurance. Unless third parties have a mechanism for inducing small firms to internalize the full liability costs *ex ante* through the purchase of full insurance, firms would find no advantage in insuring above their net worth (R. MacMinn and L. Han, 1990). Thus, legislation requiring insurance of third-party liability can be rationalized as removing the liability externality generated by small firms and levelling the playing field between large and small firms.

The CSLAA legislation also caps the liability of producers at \$500 million, above which the federal government assumes liability to third parties for the next \$1.5 billion. The effect is to formalize the externality but to transfer the external cost from potential victims to the government. The cap is redundant for firms with net worth below \$500 million but it generates a subsidy to larger firms, the value of which increases with firm size. As with the compulsory insurance provisions, the gain from caps accrues to large firms.

The preceding analysis provides an apparent rationale for the CSLAA requiring compulsory insurance. The CSLAA recognizes the apparent limits of the insurance market (coverage is seldom available for single losses in excess of \$500 million) and

¹This advantage rests on the assumption that it is costly for third parties to contract with producers. This situation may be contrasted with labor contracts where the proximity of the relationship between employee and employer, the presence of a bargaining structure and a preexisting contract relationship imply that labor can extract compensation for risk *ex ante* in the form of a risk premium in the wage.

provides a more orderly system of compensation for large losses that fall on the federal government. The Act creates a subsidy to larger firms by removing liability above the cap. While this will undoubtedly encourage higher levels of investment, whether this will lead to a more efficient allocative solution is unclear. By itself, the creation of the externality will lead to excessive investment. However, the broader policy objectives recognize the presence of substantial positive technical externalities or "spinoffs" which arise from space technologies.² Nevertheless, the benefits of the Act appear to be loaded entirely in favor of large firms. Small firms are now required to internalize the liability cost and receive no immediate benefit from the cap. Consequently, the Act apparently will deter entry of small innovative firms.

Liability Analysis with Joint Production.

The assumption that production is undertaken by a single firm is unrealistic. The technology is complex and costly, and virtually all enterprises involve many types of firms: the operator of a satellite, the launch contractor and the many subcontractors that construct the launch vehicle and its components, the satellite manufacturer and its subcontractors, as well as governments who approve and monitor the operations of the satellite. A large liability claim is unlikely to result from the negligence of a single firm but from the contributed negligence of several. Under one resolution of multiple negligence, firms may be jointly and severally (J&S) liable. Under J&S, damages may be apportioned among defendants according to the respective degrees of negligence. However, each can be fully liable for all damages. Thus should one highly culpable party become insolvent, a less culpable firm with substantial net worth may find itself fully liable for all damages. Without J&S, each firm's damages is limited by the degree of negligence it has displayed. In some jurisdictions and/or some torts, J&S operates; in others it does not.

If J&S does not operate in this setting, our analysis of CSLAA is unchanged by the introduction of joint production. If J&S does operate, it will affect contractual relationships materially. Larger solvent firms will be unwilling to commit to joint operations with small firms, unless the latter are able to internalize full liability costs *ex ante*. One way small firms may do this is by the purchase of insurance with limits in excess of net worth. Insurance may be purchased on the conventional insurance market from outside insurers. Alternatively, recognizing that large firms effectively insure the default risk of smaller coproducers under J&S, this transfer can be formalized in a hold-harmless agreement and priced accordingly. While we argued that insurance coverage in excess of net worth would not be rational under the assumption of single-firm production, this strategy is privately optimal under the assumptions of joint production and J&S liability.

The price of entering a joint venture will be higher for financially weak firms. Financially strong firms can require more risky collaborators to purchase adequate insurance. Alternatively, strong firms can supply that insurance and reflect its price in contract terms. In this way, investment in safety and loss control will be conditioned more by social costs (i.e., by expected liability) than by the wealth of the actors. With external insurance, the liability cost is internalized *ex ante* to the small firm in the insurance premium. This transfers the incentive to monitor loss control to the insurer. If the insurer is an efficient monitor, social costs will be reduced by insurance. However, given the complexity of technology, the importance of design and quality control, an *a priori* case can be made for permitting large contractors to assume the monitoring role by assuming full liability in the event of default of their collaborators.

This reasoning provides an economic rationale for J&S liability in an industry characterized by joint production. However, J&S has been criticized on allocative criteria. While it is socially desirable to internalize the liability costs jointly to those involved in production, the *ex post* allocation of liability

²See NASA 1990 annual publication *Spinoff*.

cost between the joint defendants under J&S may bear little relationship to *ex ante* behavior. Thus, J&S appears to send arbitrary cost signals. However, once we recognize freedom to contract *ex ante*, the incentives created by J&S begin to make more sense. J&S gives each firm an incentive to monitor both the loss control (safety, quality control, etc.) and the financial condition of its collaborators. Moreover, under joint production, firms with limited net worth will be required to internalize the full liability cost *ex ante* by the purchase of insurance from external suppliers or from coproducers.³

II. Conclusions

Legislation seems to be motivated by fear that uncapped liability would cause underinvestment in space enterprises and there would be a consequent failure to reap positive technical spinoffs. Liability caps counter this incentive by creating a countervailing negative externality. A second motivation is the fear that third-party victims would not have enforceable claims in tort. However, the compulsory insurance aspects of the legislation may be, at best, redundant. While liability costs appear to be externalized through the effects of limited liability, the operation of joint and several liability provides a common law alternative to compulsory insurance. Given the joint production aspects of space technology, the assumption

of liability for defaulting collaborators gives each firm an incentive to monitor coproducers and to price this default risk in joint venture contracts. Since compulsory insurance affects mostly small firms, it is likely that they would have assumed the default risk anyway *ex ante* by purchase of full insurance, or by selling this risk to collaborators in joint ventures. If insurers were better able to spread this risk and monitor safety aspects of production, then the price of entry to joint venture could be lowered by all firms fully insuring. The Act replaces this negotiable option in favor of mandated insurance. Not only is the insurance industry probably a less efficient monitor, but the infrequency of losses denies to insurers the usual mechanisms for controlling moral hazard.⁴

⁴Insurers often use experience rating to control for moral hazard when the exposure is often repeated. However, for satellite risk the sample of observations and losses is too small and thus has a high "noise-to-signal" ratio.

REFERENCES

- Kornhauser, L. and Revesz, R., "Apportioning Damages Among Potentially Insolvent Actors," *Journal of Legal Studies*, June 1990, 19, 617-51.
- MacMinn, R. and Han, L., "Limited Liability, Liability Insurance and Corporate Risk Management," Working Paper, University of Texas, 1990.
- Ringleb, A. and Wiggins, S., "Liability and Large Scale, Long-Term Hazards," *Journal of Political Economy*, June 1990, 98, 574-95.
- NASA Scientific and Technical Information Facility, *Spinoff*, 1990.

³L. Kornhauser and R. Revesz (1990) have recently argued that, with joint tortfeasors and the potential for insolvency, one cannot make a general statement about the efficiency aspects of J&S. Our stronger statements about the efficiency effects of J&S relate to the current narrowly defined problem. We do not imply here that they can be generalized.

The National Aerospace Plane: An American Technological Long Shot, Japanese Style

By LINDA R. COHEN, SUSAN A. EDELMAN, AND ROGER G. NOLL*

In the mid-1980's, the United States embarked on a most ambitious government research and development (R&D) venture, the National Aerospace Plane (NASP). NASP's goal is to build a piloted research plane, the X-30, that will travel 25 times the speed of sound, use normal airports, and fly to near-earth orbit. Moreover, the program's organization has a distinctly Japanese form: five government agencies and five companies that might have competed for the program are collaborating on a single design.

This paper examines the match between NASP's technical challenges and its political and organizational incentives. We conclude that the technical basis for NASP is shaky, and that its organizational innovation increases the chance of failure.

I. The Political Economics of R&D Programs

The positive economic theory of government R&D programs is based on three results from the rational actor model of politics (see Cohen and Noll, 1991). First, voters evaluate incumbent politicians retrospectively (what have politicians done for us lately?), causing incumbents to favor programs with immediate payoffs. Second, because citizens largely organize political participation around groups, political actors favor programs that deliver visible rewards to organized interests. Hence program expenditures are often political benefits rather than costs. Moreover, when dealing with an industry, political actors prefer a program that distributes benefits among all industry participants, rather than one that picks winners or otherwise disadvantages some firms. Third, because incumbent politicians have

high probabilities of reelection, they fear political mistakes more than they value political windfalls. Hence, they seek to avoid programs with risky benefits.

All of these characteristics make R&D politically unattractive. R&D usually has a long time horizon with uncertain payoffs, and can upset wealth positions in industry. If the most efficient structure of an R&D program is to centralize management, the prime contractor may gain a competitive edge. Finally, the distributive benefits from contract expenditures are smaller in the early research phase, so that political leaders have an incentive to speed up research in order to get on with the big-ticket prototype and demonstration phases. Insufficient research lowers the expected net benefits of a program.

Industrywide collaborative ventures deal with some of the political problems of R&D projects. Most clearly, they minimize the disruptive act of "picking winners." Because collaboration lowers the likelihood of independent R&D by each collaborator, it reduces the chance that the project will fail because someone else invents something better. Nevertheless, collaborative government R&D ventures have been rare. The Clinch River Breeder Reactor involved a large number of utilities, and Sematech includes most of the semiconductor industry. But other ventures, notably other commercial aerospace ventures (satellites, supersonic transport, Space Shuttle) used competition among several contractors. In the case of satellites, steadfast pursuit of a procompetitive policy caused an otherwise successful program to be cancelled (Cohen and Noll, ch. 7).

Collaborative ventures have two undesirable characteristics. They narrow the range of alternative paths of development that are explored and can cartelize the industry. The first becomes a political cost if the collaborators produce no worthwhile advance, but

*University of California, Irvine, CA 92717; Columbia University, New York, NY 10027; Stanford University, Stanford, CA 94305, respectively.

other promising paths go unexplored or, even worse, are developed by foreigners. The second is politically costly if cartelization becomes a political issue, such as might occur if the collaborators sell to another politically influential industry. In the past, political actors have eschewed the collaborative approach. However, the principle has become fashionable in government, and is a major component of the recent technology policy plan issued by the White House, which calls for antitrust exemptions and other policies to promote R&D joint ventures in response to perceived Japanese successes on "critical technology" fronts (Office of Science and Technology Policy, 1990).

II. The NASP Program

The NASP program dates to 1984, when the Defense Advanced Research Projects Agency (DARPA) became convinced that hypersonic technology was worth pursuing (T. A. Heppenheimer 1987). During 1985, four other agencies were brought into the program: Air Force, Navy, National Aeronautics and Space Administration (NASA), and the Space Defense Initiative Organization (SDIO).

Two technical advances made DARPA enthusiastic about the program. One was in scramjets—air-breathing engines capable of hypersonic flight. Air-breathing engines, unlike rockets, need not carry oxygen for fuel combustion, and so reduce weight at take-off. The second advance was in materials. Very high speeds cause an airplane to become extremely hot. The challenge is to create strong, heat-resistant, lightweight materials.

System integration issues also present significant challenges. To exceed speeds of about Mach-7, the plane must use hydrogen slush as a fuel. This requires building tanks and pipes for storing and delivering hydrogen to the engines. Moreover, the slush will be used to cool the craft. Constructing the fuel and cooling systems is not trivial. Another system integration issue concerns airframe design. Scramjet engines must be extremely long, or else the air (traveling at Mach-25) and fuel will not combust inside the engine. Engines are part of the airframe

and affect airflows (and lift and drag); very large engines present an unsolved fuselage design problem. Finally, the program places heavy reliance on computer simulations. Current wind tunnels simulate conditions only up to Mach-7; to test the plane requires actual flight. Flight simulations require unprecedented advances in computer fluid dynamics.

In sum, NASP leaps into uncharted technical waters. Mach-25 is more than three times faster than the existing speed record for experimental aircraft and seven times faster than any aircraft has flown in sustained flight. Every important element of the program attempts a radical advance of known technology.

The NASP management is also innovative. Initially, five aircraft and engine firms independently pursued designs. In May 1990, the groups working on NASP within each firm were merged into a single "team" that shares all technical results and is formulating a single design incorporating features of each firm's past work. The team members have cost-plus-award contracts with the government, and will share the award (profit) equally, regardless of the ultimate division of work ("Teaming Agreement Stresses Equality Among Five NASP Prime Contractors," 1990).

The management form increases the riskiness of an already risky project. Exactly one engine, airframe, and fuel system will be developed for a radical technology. Moreover, the incentives for each contractor to devote best efforts to the project are diminished by severing the tie between technical contribution and shares in profits and information.

III. The Politics of NASP

To understand the politics of NASP requires identifying the reasons each major player participates in the program. The aircraft industry receives several hundred million dollars per year for research on advanced engine and airframe design, and will get substantial contracts once construction of the X-30 begins. Industry will benefit regardless of whether the X-30 reaches orbit or even flies.

The rationale for government is more complex. Because of the high costs and risks of the program, none of the agencies involved has been willing to commit to it to satisfy policy missions. The committee approach, however, has been (marginally) acceptable to the agencies; more importantly, it created a broad coalition favoring NASP within Congress and the White House.

The Air Force and Navy entered the program during the era of the "evil Empire," and so were motivated by the arms race. The end of the Cold War reduced the budget for defense R&D. Although reversed by both the White House and Congress, the Air Force attempted to pull out of the program in 1989, proposing to reallocate the budget to other defense programs (Craig Covault, 1989; U.S. House of Representatives, 1989).

NASA, whose top priority is piloted space flight, has a far greater stake in the program. The agency's present technology (the Space Shuttle) borders on fiasco. However, NASA steadfastly maintains that NASP is a research venture without an explicit mission and not a substitute for other advanced launch technologies. SDIO currently is pursuing a separate single-stage-to-orbit vehicle (James Asker, 1990). DARPA was too small from the start to continue its more limited program in scramjets.

Coalitional support means the program must demonstrate technology that relates to the missions of the different agencies. Follow-on military applications require sustained hypersonic flight; space applications must attain orbit. The latter requires the X-30 to fly not just hypersonic, but at Mach-25; additionally, the current design calls for a rocket to be incorporated into the aircraft for the final thrust to orbit. Satisfying NASA's congressional overseers requires the X-30 to be piloted by people (U.S. House of Representatives, pp. 51-52).

Our analysis suggests a paradoxical relationship between agency budgets and project scope. When agencies are rich, they can pursue small projects with narrow objectives. For example, NASA's initial foray into geosynchronous satellite technology occurred in the early 1960's when government enthusiasm for space-related activities sur-

passed NASA's ability to spend. Alternatively, poor agencies need to mobilize large coalitions and consequently pursue ambitious programs with multiple objectives. The government team approach for NASP increases the technical complexity of the program, and significantly raises risks of failure along any dimension.

The NASP program's management strategy, consistent with efficient sequential decision making for R&D, is to postpone the decision to build the X-30 until the design phase is successfully completed. Unfortunately, the government cannot commit to cancel the program should the design phase produce bad news. NASP's two most recent aerospace predecessors provide sobering examples. The initial supersonic transport (SST) design proved infeasible, and the plane's performance characteristics never lived up to initial goals (Edelman, 1991). The SST had the same management plan as NASP, but was not killed until 1971, not because of performance shortfalls (which had been known for 5 years), but because of opposition from environmentalists at the peak of their influence in Congress. Similarly, the component that made the Space Shuttle's *ex ante* expected net benefits positive (the Space Tug, a vehicle carried in the Shuttle's bay that could fly to geosynchronous orbit) proved technically infeasible prior to construction of the shuttle fleet (Jeffrey Banks, 1991).

Whether NASP can be cancelled if the design results are discouraging depends on the incentives facing Congress and the president. Both are likely to focus on aspects of the program of less concern to the agencies. One is the symbolic value of the technology. The other is the political value of NASP contracts. Our work on other large R&D programs finds that political support for continuing failed projects is significantly affected by the magnitude of current expenditures.

The symbolic value of the program has found continued expression among political supporters. President Reagan announced the program in his 1986 State of the Union address, when he spoke not of the need to support R&D in the aircraft industry but instead of the "Orient Express" of the

twenty-first century—a plane that would allow passengers to fly from New York to Tokyo in an hour or two. Congress has focused on specific applications, such as replacing the Space Shuttle or building a reconnaissance plane as fast as a missile despite the fact that these objectives could be achieved more cheaply and with greater certainty by other technologies (Stephen Korthals-Altes, 1987; Bruno Augenstein and Elwyn Harris, 1989; Office of Technology Assessment, 1989). Industry regarded the team approach as increasing the political feasibility. According to Barry Waldman, leader of the industry team, “If we don’t pull together, we don’t get to Phase 3,” the high-budget construction of X-30s (“Teaming Agreement...,” p. 38). Attempts to stress that the program is pure R&D are inconsistent with the basis for its political support, and so are probably irrelevant. To cancel the program after the design stage would not only harm the contracting companies, but would prevent attaining a symbolic accomplishment that political leaders could use to their benefit.

IV. Conclusions

It is unlikely that NASP will be more than the agencies contend: an exciting research venture that will produce new aviation technology and, if the X-30 flies, set new speed records. Nonetheless, the program has been designed to produce another technological turkey reminiscent of past big-ticket failures. The innovative “team” approach makes the program more difficult to kill if NASP becomes nothing more than an expensive toy. Involving all of the important players in the aerospace industry eliminates short-term sources of political attack because it picks no winners and has no competitive external R&D effort. Involving multiple government agencies creates a stable support coalition within government.

The NASP program’s organizational structure is born of political forces, not technical and economic optimization. High-risk projects call for incremental development, flexible management schedules and the ability to explore many technical paths before committing to a system design.

Thus, the “Japanese-style” government-industry collaboration serves political needs at the expense of expected program performance.

REFERENCES

- Asker, James, “SDI Organization Plans 1994 Test Flight of Single-Stage-to-Orbit Spacecraft,” *Aviation Week and Space Technology*, November 5, 1990, 133, No. 19, 26–27.
- Augenstein, Bruno and Harris, Elwyn, “Assessment of NASP: Future Options,” Working Draft WD-4437-AF, Rand Corporation, June 1989.
- Banks, Jeffrey, “The Space Shuttle,” in L. Cohen and R. Noll, eds., *The Technology Pork Barrel*, Washington: Brookings Institution, 1991.
- Cohen, Linda and Noll, Roger, *The Technology Pork Barrel*, Washington: Brookings Institution, 1991.
- Covault, Craig, “White House Acts to Reverse Aero-Space Plane Cancellation,” *Aviation Week and Space Technology*, April 24, 1989, 130, No. 17, 20–21.
- Edelman, Susan, “The American SST,” in L. Cohen and R. Noll, eds., *The Technology Pork Barrel*, Washington: Brookings Institution, 1991.
- Heppenheimer, T. A., *The National Aerospace Plane*, Arlington: Pasha Publications, 1987.
- Korthals-Altes, Stephen W., “Will the Aerospace Plane Work?,” *Technology Review*, January 1987, 90, No. 1, 43–51.
- Office of Science and Technology Policy, U.S. *Technology Policy*, Washington, D.C., September 26, 1990.
- “Teaming Agreement Stresses Equality Among Five NASP Prime Contractors,” *Aviation Week and Space Technology*, October 29, 1990, 133, No. 18, 38–39.
- U.S. House of Representatives, *The National Aerospace Plane*, Joint Hearing No. 53, Committee on Science, Space, and Technology, 101st Congress, 1st Session, August 2, 1989, Washington: USGPO, 1989.
- U.S. Office of Technology Assessment, *Round Trip to Orbit: Human Spaceflight Alternatives—Special Report*, Washington: USGPO, 1989.

TORT LAW AS A REGULATORY SYSTEM[†]

Regulation and the Law of Torts

By SUSAN ROSE-ACKERMAN*

Tort law is "private" law. Regulation by statute is "public" law. How should the two relate to each other in a regulatory state where statutory intervention in private markets is widespread? Both tort and statutory law have regulatory effects. Thus economic policymakers should examine the links between these legal regimes in substantive areas where both systems operate. I argue in this paper that statutes should generally dominate so long as agencies can use rule-making to shape policy. Common-law torts should be limited to areas of activity not covered by statutes and to situations in which courts can complement the statutory scheme with a supplementary enforcement and compensation mechanism. However, if the implementation of a statutory scheme requires that private tort actions be preempted, alternative methods of compensating victims, such as social insurance or statute-based private damage actions, may need to be designed.

I. Torts vs. Statutes

The fundamental differences between tort law and regulation center not on substantive standards, or on the distribution of benefits and harms, but on procedures. Statutory regulation, unlike tort law, uses agency officials to decide individual cases instead of judges and juries; resolves some generic issues in rulemakings not linked to individual cases; uses nonjudicialized procedures to

evaluate technocratic information; affects behavior *ex ante* without waiting for harm to occur, and minimizes the inconsistent and unequal coverage arising from individual adjudication. In short, the differences involve who decides, at what time, with what information, under what procedures, and with what scope.

Steven Shavell has developed a useful four-category schema to organize a discussion of alternatives (1987, pp. 277-90). He distinguishes between *ex post* (backward-looking) and *ex ante* (forward-looking) options, and between privately initiated and state-initiated systems. This framework produces four alternatives: Tort liability (*ex post*, privately initiated); court injunctions (*ex ante*, privately initiated); command-and-control regulation or corrective taxes (*ex ante*, state initiated); and fines for harm done (*ex post*, state initiated). For our purposes, the most important comparisons are between tort liability and *ex ante*, state-initiated approaches.

Five factors, according to Shavell, should influence the choice between *ex ante*, state-initiated and *ex post*, privately initiated approaches. First, state action is desirable when the harm is so diffuse that individuals have little incentive to sue on their own and cannot cheaply organize to sue as a group. Second, if injurers are too poor to pay for the harm they cause, a system based on *ex post* payments will not effectively deter them. Third, when harm can be demonstrated on a statistical, but not an individual, basis, regulations or taxes applied *ex ante* can shape behavior without a showing of causal links between particular parties.

Fourth, an *ex ante* regulatory system will be preferable when the same information about costs and benefits is relevant to many

[†]*Discussants:* Roger Noll, Stanford University; George Eads, General Motors Corporation; Robert Crandall, The Brookings Institution.

*Ely Professor of Law and Political Economy, Yale University. This paper is adopted from my forthcoming article (1991).

instances of harm. Fifth, administrative costs are an important consideration. If the probability of harm is low, *ex post* systems may be preferable since they only need come into play when damage occurs.

Ironically, tort law may be most ineffective in precisely those areas where judicial doctrine has been most innovative—toxic torts, products liability, and medical malpractice. These are all areas where *ex ante* regulation enjoys distinct advantages. Some critics, such as Peter Huber (1988) and W. Kip Viscusi (1984), point to the courts' incompetence concerning technical issues of health and safety—incompetence arising from lack of expertise, inadequate staff, and procedures ill-suited to the discovery of scientific truth.

Nevertheless, widespread support for the tort system persists even when the logic of efficient risk control demands *ex ante* regulation. Given this reality, the remainder of this paper attempts to explain how torts can be used to further, not distort, deterrence goals and isolates those situations where preemption would be desirable.

II. Torts as a Complement to Statutory Regulation

Let us consider, then, a situation favoring statutory regulation over torts—for example, the safety of automobiles, drugs, or medical devices, where many people are at risk and information about product design is often relevant to harm. To highlight the differences between statutes and torts, assume that regulatory standards are set through rulemaking under a command-and-control scheme.

Ideally, tort law and regulatory standards work together to further deterrence and compensation goals. Torts and regulations can be complementary: 1) when tort doctrines are stopgaps which apply absent more stringent statutes; 2) when regulatory standards are intended as minima which more stringent tort doctrines can supplement, and 3) when a regulatory standard is set at the socially optimal level and tort doctrine imposes either strict liability or a standard of care lower than that required by the agency.

However, when the regulatory standard is set optimally while the tort standard of negligence is interpreted to require an even higher level of care, conflicts can arise under a system of compensatory damages. Such conflicts are even more likely when punitive damages are available.

Legislatures have other aims besides constructing rational regulatory systems, and they often fail to establish regulatory programs in all of the areas in which regulation would in theory be superior to tort law. In such cases the first type of complementarity applies. Courts should not idealize the pattern of common-law regulation created by their past adjudications, but should instead see tort law as a stopgap pending future statutory regulation. "The common law standard of [reasonableness]...can at least serve the needs of our society until the legislature imposes higher standards" (*Larsen v. General Motors*, 1968, p. 506). If a regulatory statute is then passed, courts should resolve conflicts between tort doctrines and regulatory principles by according priority to the statute (*Wood v. General Motors Corp.*, 1989, p. 402).

Regulatory standards are sometimes designed to establish only a baseline. Then the second case holds. While violation of such a standard usually amounts to negligence per se in a tort suit, compliance with the standard is merely evidence for the jury to consider in determining reasonable conduct. Note the asymmetry here. Because the standard sets a minimum, the plaintiff can argue that a higher standard should be imposed in a particular case, but the defendant cannot invoke special circumstances to justify its violation of the basic standard.

The courts have taken this approach in product and occupational safety cases; they have ruled, for example, that the Food, Drug and Cosmetic Act sets only minima and does not preempt tort suits for damages. Similarly, tort suits involving automobile design or exposure to nuclear materials (*Silkwood v. Kerr-McGee*, 1984) are not preempted by regulatory statutes.

These first two ways of reconciling tort law and statutory law are logically inconsistent but coexist in practice; tort suits with

regulatory effect are permitted both when the legislature has not acted and when it has established a statutory minimum. Yet regulations cannot really be minima if the common law operates as a stopgap.

If the courts view regulatory standards as minima, there is no conflict with the statutory scheme if the agency itself has set a low standard in the belief that case-by-case adjudication is the best way to respond to the regulatory problem. Although this may sometimes be a plausible strategy, the plaintiff should bear the burden of demonstrating that the legislature intended it. Absent such a showing, a regulatory statute should be taken to imply a legislative judgment that a comprehensive, state-centered, *ex ante* approach is the best way to deter harm. The difficulties of judicial standard setting should lead judges to accept the stopgap role. If the regulations are indeed too lax, statutory or administrative reform is appropriate, not *ad hoc* judicial actions that respond to individual needs while producing systemwide inequities and inefficiencies.

Under the third form of complementarity, the doctrines of either negligence or strict liability can produce both a more consistent tort law and a more effective regulatory system. To accomplish this, however, courts must be prepared to surrender some of their independence in setting standards of care and assessing damages.

Consider a negligence rule that seeks to mimic the regulatory standard. In contrast to the asymmetric doctrine of negligence *per se*, the courts would also recognize a *per se* defense for injurers who meet the regulatory standard. The wrongdoer could thus be punished twice: by whatever sanctions the state imposes through the regulatory process, and by paying damages to private litigants. While this may seem unfair to the wrongdoer, it is not inefficient even if the sum of the penalties exceeds the social costs of violating the standards. Convicted wrongdoers would pay "too much," but anyone can avoid this overcharge by simply conforming to the regulatory requirements. This optimistic view is true, however, only if agencies set clear standards, courts accept these standards in determining liability, and apply them competently to individual cases.

Suppose now that the tort standard is not negligence but true strict liability, which holds a manufacturer liable for all harm caused by its product and only requires the court to determine causation, not to assess the risks and benefits of product design. True strict liability differs substantially from the "strict liability" of products law, which essentially requires the jury to make a negligence-like risk/utility calculation. Under true strict liability, torts and regulation need not conflict if damages are set equal to the harm caused by the tortfeasor. The possibility of a tort judgment will simply give the regulated entity an additional incentive to comply with the statute. The conditions assuring complementarity, however, do not now exist. To achieve them would require substantial tort reform, yet attempts to reshape tort doctrine according to its behavioral effects would be difficult.

In contrast to the complementary options outlined above, tort law can work at cross purposes to statutes when the regulatory standard is set at the socially optimal level of care, but the courts impose a more stringent negligence standard. Two cases need to be considered: compensatory damages and punitive damages.

Judges sometimes sharply distinguish the two, finding punitive damages "regulatory" but not compensatory damages (for example *Silkwood v. Kerr-McGee*, Powell dissenting, p. 276). Under this view, the goal of compensatory damages is merely to make the victim whole, not to induce behavioral changes in potential injurers; thus compensatory damages cannot conflict with a regulatory purpose. If tort actions only provided compensation and if the agency's own enforcement mechanisms effectively assured compliance with its standards, this view might be correct. But these assumptions are always false where no statute exists (the stopgap case) and are often false even when one does.

To see the judicial error, suppose that a regulatory agency has set a product standard at the optimal level, but courts nonetheless find that complying firms have been negligent or have defectively designed their product. A firm will then compare the extra costs of meeting the court's standard

with the damages it must pay victims if it merely complies with the lower regulatory agency standard. If it can increase profits by complying with the court's negligence standard, the firm will do so. At this care level, however, it will surpass the optimal agency standard, and marginal costs will exceed marginal benefits.

Judges who distinguish between compensatory and punitive damages, however, are not completely misguided. Punitive damages also influence caretaking and product design but may produce a different outcome than compensatory damages. If the courts impose a higher standard of care than is required by socially optimal statute, economic distortions will occur. The regulated firm subject to punitive damages will choose one of two options (depending upon which is profit maximizing): compliance with the tort standard where no damages are levied, or a level of safe design (somewhere between the socially optimal level and the tort standard) at which its marginal costs equal marginal punitive damages.¹

III. Regulatory Reform

Commentators have long urged legislators and regulatory agencies to charge fees set to reflect the risks created by regulated firms. Incentive-based reforms allocate regulatory costs to those who can bear them most efficiently, encourage firms to search for innovative ways to reduce harms, and force producers to reflect the risks they impose on society.

Such reforms, however, could be undermined by a poorly informed judiciary. If courts equate regulation with standard setting, then they may treat only command-and-control regulation as behaviorally significant. In a recent case, for example, the Superfund law was described as "not a regulatory standard-setting statute" because polluters pay for the cost of abating hazardous wastes "through tax and reimburse-

ment liability" (*State of New York v. Shore Realty*, 1985, p. 1041).

Incentive schemes require a fundamental rethinking of the relationship between tort law and statutory law. How should courts handle claims by defendants that incentive-based regulatory statutes preempt tort actions? Judges who view regulation as confined to standard setting might allow tort actions on the ground that these statutes are not "regulatory" because they do not establish uniform standards but "only" create incentives. Yet the argument for preemption of tort law is even stronger in the case of incentive-based regulations than in the case of command-and-control regulation. A well-designed incentive system signals to a firm the social costs of its activities. A fee system resembles a tort liability system: No fixed standards are set, but firms respond to the cost of damages. The regulated entity must purchase the right to impose social costs in the same way that a tort judgment requires payment for harms. The main difference is the comprehensiveness of a fee schedule, which the state sets so that all firms are covered. A firm's liability does not depend on the contingency of private litigation and jury damage awards.

If fee schedules are set to reflect the social costs of the regulated firm's activities, tort judgments would undermine the regulatory scheme, especially if courts applied a strict liability standard, the type of standard which some judges have found least "regulatory" (*Silkwood v. Kerr-McGee*, p. 276 n. 3). Thus incentive-based statutes should include a provision clearly preempting tort actions. The only role for lawsuits by private individuals would be to force the *agency* to enforce its own rules; such suits might permit private recovery of damages for harm caused by lax enforcement. Although ordinary tort actions would be preempted, certain specialized private remedies might supplement agency enforcement just as tort actions do which use regulatory standards as the standard of negligence.

IV. Compensation

Tort law provides more than a set of regulatory incentives; behavior modification

¹ Punitive damages may reflect not punishment of the tortfeasor but a correction for the fact that only a fraction of victims sue. If the courts set the damages multiplier to correct for this, the earlier analysis of compensatory damages applies.

is not its only legitimate function. It is also a compensation system triggered by victims' complaints. If a regulatory statute bars private tort actions, those who were previously able to sue for damages will be disadvantaged, a result courts seem reluctant to permit (*Silkwood v. Kerr-McGee*, p. 251). If compensation of victims is not addressed by a purely regulatory statute but remains a policy goal, courts may apply conventional tort doctrines that are at cross purposes with regulatory policies. Yet retaining conventional tort actions in the face of regulatory statutes can undermine the behavioral impact of statutes. Other solutions must be found to the problem of providing compensation.

If the victims are numerous and their losses fall into broad, easily identified categories, such as lost limbs or particular types of cancers, then the compensation goal could be served by direct subsidy programs like workers' compensation and the black lung program. In contrast, if the victims are few in number and their problems are idiosyncratic, the law should either permit private rights of action for damages analogous to those permitted under the Consumer Product Safety Act and the Comprehensive Environmental Response, Compensation and Liability Act (Superfund),² or it should allow tort actions under strict liability principles solely as a means of achieving compensation.

V. Conclusions

The tort system has shown itself inadequate to deal effectively with problems, such as latent cancer risks and attenuated chains of causation, which do not fit easily into

traditional tort categories. The innovations that the courts have developed to manage class actions and consolidate cases are transforming the courts into quasi-regulatory agencies. Real agencies are likely to perform better than awkward judicial hybrids that have many of the disadvantages of both forms.

Reform-minded policymakers should carefully reexamine the relationship between tort law and statutory regulation. The widespread presumption favoring a vigorous tort system should be replaced with a more comprehensive view of the alternative ways to achieve both deterrence and compensation. Privately initiated lawsuits, brought either under tort principles or under regulatory statutes, should be quite limited and targeted on augmenting regulatory enforcement and responding to unusual situations that would be poorly resolved by broad-based regulations.

REFERENCES

- Huber, Peter, *Liability: The Legal Revolution and Its Consequences*, New York: Basic Books, 1988.
- Rose-Ackerman, Susan, "Tort Law in the Regulatory State," in P. Schuck, ed., *Tort Law and the Public Interest: Competition, Innovation, and Consumer Welfare*, New York: Norton, forthcoming 1991.
- Shavell, Steven, *Economic Analysis of Accident Law*, Cambridge: Harvard University Press, 1987.
- Viscusi, W. Kip, "Structuring an Effective Occupational Disease Policy: Victim Compensation and Risk Regulation," *Yale Journal on Regulation*, 1984, 2, 53-81.
- Larsen v. General Motors*, 391 F. 2d 495 (8th Cir. 1968).
- Silkwood v. Kerr-McGee*, 464 U.S. 238 (1984).
- State of New York v. Shore Realty*, 759 F. 2d 1032 (2d Cir. 1985).
- Wood v. General Motors Corp.*, 865 F. 2d 395 (1st Cir. 1989).

²Under the CERCLA, private individuals can sue generators of hazardous wastes for cleanup costs (but not for personal injuries) even if the government has taken no action against the waste generator.

The Safety and Innovation Effects of U.S. Liability Law: The Evidence

By ROBERT E. LITAN*

Although it has vanished from the front pages, the "liability crisis" still continues to occupy the attention of policymakers, the business community, and consumers. Indeed, many states have responded to the crisis by enacting "tort reforms" to make it more difficult for accident victims to sue and, if they win, to recover large awards.

Manufacturers and insurers have supported these efforts because they believe that the tort system has become so expensive and uncertain that is "overdetering," discouraging the production of valuable products and the innovation of many more. Plaintiffs' lawyers and consumers groups vigorously disagree, arguing that current liability law provides important incentives for private actors to make safer products and to conduct their activities in a less risky fashion.

Given the significant impacts of the tort system (by one recent estimate, gross liability expenditures in the United States amounted to \$117 billion in 1987, see Tillinghast, 1989), it is somewhat surprising that there is so little empirical research to settle this debate.

In fact, there even is a lack of consensus on liability trends. Although there are no comprehensive nationwide data on the numbers of nonautomobile tort suits, verdicts, and settlements, by all available measures, total liability costs (primarily insurance premiums, but also estimates of self-insurance payouts) have been rising during the last 30 years, and especially rapidly in the 1980's (U.S. Department of Justice, 1986; Mark Peterson; 1987).

In 1989, the Brookings Institution launched a major study designed to determine the "safety" and "innovation" impacts of this expansion in the tort system on five sectors of the U.S. economy; the private aircraft, automobile, chemical and pharmaceutical industries, and the medical profession. As project directors, Peter Huber and I commissioned experts to address each side of the debate for each of these sectors, deliberately seeking individuals who were *not* economists. As discussed below, it is extremely difficult (some would say impossible) to establish through standard econometric techniques the linkages, if any, between liability law and safety and/or innovation. Accordingly, we believed the project could shed more light on the subject by engaging individuals with scientific or practical backgrounds in the specific sectors they were asked to study; in a few cases, economists met these criteria and they participated. This article summarizes some of the key findings from this effort (1991).

I. A Brief Review of the Preexisting Literature

The "evidence" that liability has had a "safety-enhancing" effect is largely anecdotal. Recent books and articles, for example, have pointed to specific substances and devices that allegedly or actually have led to mass harms (Agent Orange, the Dalkon Shield contraceptive, asbestos, and tampons) and subsequently have been removed from the marketplace in the aftermath of tort litigation. In addition, there are many other accounts of unsafe products or practices that have been attacked, successfully or otherwise, in court, but it is often difficult to determine from these accounts whether tort verdicts were the reason for the remedial actions that supposedly followed the litigation, or alternatively that these remedies cured real problems. Similarly, while

*Senior Fellow and Director, Center for Economic Progress and Employment, Economic Studies Program, The Brookings Institution, 1775 Massachusetts Avenue NW, Washington, D.C. 20036.

George Eads and Peter Reuter (1983) find from their interviews of corporate managers that product liability exerts a strong "pro-safety" effect on product design, they also confess that current liability law sends an "extremely vague signal" since it does not indicate "*how* to be careful, or more important, *how* careful to be" (pp. viii-ix).

Statistical analysis of the safety effects is rarer, but the little that exists reports little impact. Thus, George Priest (1989) finds that while the annual numbers of tort suits and aggregate liability insurance premiums rose sharply during the 1980's, injury rates for consumers and workers, death rates from medical procedures, and aviation accident rates showed no more rapid a decline in that decade than in the 1970's when the volumes of tort suits and premium costs were far lower. Obviously, a demonstration of this sort does not distinguish the contributions that other factors (notably, consumer demand for safety and/or preventative measures taken by consumers, and regulation) might have made to the declining injury and death rates. But Priest's work nevertheless raises doubts that the safety benefits from expanded liability have been large.

The literature on the innovation side of the debate also has been relatively sparse, and to some extent contradictory. The Conference Board, for example, has put out two studies with conflicting conclusions: an early survey of corporate risk managers who report relatively little harmful effect of liability (E. Patrick McGuire, 1988) and a follow-up survey of chief executive officers who report much more substantial negative effects (Nathan Weber, 1989). There are also anecdotes of arguably socially useful products and services (principally medical) withdrawn from the market allegedly because of the uncertainty created by the liability climate.

Economists will have a hard time (to put it mildly) using traditional statistical methods sorting out the conflicting claims about tort law's impacts in a convincing fashion. First, it is difficult to specify and/or gather data on the appropriate dependent variable. Although safety is relatively easy to mea-

sure (by such variables as injury and death rates), objections may be lodged against all of the standard measures of innovation: by inputs (such as R&D expenditures) or outputs (such as the numbers of patents or new products introduced).

Second, and more troubling, is the difficulty of specifying and obtaining adequate data representing the influence of liability. As noted earlier, there are no comprehensive time-series data for a sufficient period of time to perform meaningful statistical analysis on the total numbers of tort suits (in both federal and state courts combined), favorable verdicts and/or settlements, or dollars paid through or on account of the liability system, especially by industry.

Third, any worthwhile statistical analysis of liability's impact must account for the influence of other important variables—perhaps, most importantly, regulation. Yet how does one reliably measure regulatory intensity? Proxies such as the numbers of regulatory personnel and dollars spent by the agencies that employ them ignore important distinctions between the way regulation is implemented (through performance standards or design controls, for example) and thus may not capture regulation's true effects.

Notwithstanding these difficulties, it may still be possible to make rough estimates of liability's effects during the limited time periods for which data are available. But with the relevant statistical data so thin, it is useful in the meantime to collect information through other means.

II. Safety Impact¹

Most of the authors who addressed the safety side of the liability debate found little direct or statistical evidence that specific liability verdicts have led to substantially safer products. To be sure, tort suits often induce manufacturers to expand or modify warnings, but there also is little concrete

¹Authors cited in Sections II and III can be found in my 1991 study with Huber.

evidence that such redesigned warnings have resulted in fewer deaths and injuries.

This does not mean that the authors found that liability has no safety effects at all. Instead, there is evidence that tort law has a safety-enhancing impact that works *indirectly* through the adverse publicity that tort suits, and more importantly, proplaintiff verdicts generate about particular products. Thus, for example, John Graham demonstrates that liability verdicts involving the Ford Pinto and the "CJ" Jeep led to negative press stories (especially on national television) that in turn diminished consumer demand for these vehicles.

But when it comes to product design, our authors for the most part concluded that the most important impulses for safety come from *outside* the liability system: through a combination of consumer demand for safety, standards for professional liability (in the case of medical services), and regulation (especially in the case of automobiles, pharmaceuticals and private aircraft). For products with long lives (such as aircraft and even cars) this is not surprising. Product litigation in such cases often arrives years after designs have changed for market-driven reasons or in response to prodding by regulators.

Our authors also found evidence supporting the view expressed by Eads and Reuter that the liability system sends confusing signals to private actors, in significant part because courts and regulators often have reached markedly different conclusions about how accidents are caused. This, too, would help account for the weak effects of liability on safer design.

Louis Lasagna, for example, points to the dramatic differences between the favorable analysis and pronouncements of the Food and Drug Administration on the dangers of Benedictin (a drug for preventing nausea associated with pregnancy) and various vaccines, and the far more negative verdicts on the same products rendered by the legal system. Similarly, Robert Martin describes a sharp divergence between the postaccident findings of the National Transportation Safety Board and the Federal Aviation Administration (which during the sample pe-

riod he analyzed found that design or manufacturing defects accounted for *none* of the more than 200 accidents) and the many lawsuits filed in connection with the same accidents claiming that defects were the sole cause of the mishaps. Finally, Stanley Reiser suggests that because few doctors believe the liability system produces accurate results, doctors have altered their practices in ways that are far more costly (due to the practice of "defensive medicine") and potentially less conducive to patient welfare (due to the erosion of trust formerly inherent in the doctor-patient relationship).

Indeed, several of our authors maintained that the threat of product liability can *deter* safety improvements. One automobile manufacturer feared making design changes that could easily be interpreted by juries as an admission that the prior design was defective. In one area where changes *were* made (the decision by automobile manufacturers to phase out tension relievers in seat belts, in part for liability reasons), Graham points out that the replacement "slack" seatbelts are less safe than their predecessor. Similarly, Martin found that in comparison to the prior decades when liability litigation was not that pervasive in the aircraft industry, the rate of decline in the aircraft accident rate *slowed* during the 1970-90 period when tort litigation was much more intensive. From this, Martin concluded that recent liability trends actually may have retarded the introduction of safer designs (although Martin's statistical analysis did not control for other factors that may have affected trends in aircraft accident rates).

A key exception to these relatively negative verdicts on the safety impacts of the current tort system is a paper by Nicholas Ashford and Robert Stone, which argues that at least in the chemical industry (and perhaps elsewhere), liability law *underdet*ers. Ashford and Stone contend this is largely because it is extremely difficult for juries and courts to link specific chemical agents with particular adverse health outcomes (except for such "signature" diseases as asbestosis and mesothelioma in the case of asbestos, for example). In addition, the victims of harmful chemicals are often

workers, whose compensation for noneconomic damages is strictly limited by workmen's compensation laws.

Yet even Ashford and Stone must concede that workers can still file tort claims against the *manufacturers* of harmful substances and in this way circumvent the restrictions of workmen's compensation, which of course, is exactly what asbestos victims have done. Meanwhile, while it may be true *ex post* that liability may underdeter where science has not yet established cause-and-effect relationships between particular substances and human health, this does not mean that *ex ante* liability fails to deter. To carry the weight of their argument, Ashford and Stone must demonstrate that juries and courts consistently ignore causal linkages that the scientific community has already demonstrated. While they present suggestive evidence that the risks of certain chemicals *may* be more dangerous than certain government agencies and courts have admitted, it is this author's view that the jury is still out on these claims.

III. Innovation Impact

Unlike the attempts to determine the safety impact of liability, which at least in principle involve observable events (a verdict and subsequent action), efforts to isolate the effect of tort law on innovation must uncover the proverbial "dogs that don't bark," or events that *do not* happen. Given the difficulty of this challenge, the Brookings researchers generally drew inferences from liability-related events that have occurred, much like physicists who infer the existence of subatomic events from the "traces" such particles leave behind.

Thus, Martin makes the case that the combined effects of uncertainty and high awards and settlements have largely decimated the private aircraft industry (where annual production levels have fallen by more than 80 percent from pre-1970's peaks), while contributing to the declining rate of introduction of new light aircraft. He estimates that liability costs added an average of \$70,000–\$100,000 to the costs of private airplanes built in 1987.

Lasagna finds that liability has discouraged research into certain types of pharmaceuticals and vaccines (especially those that could benefit children and pregnant women).

Murray MacKay argues that fears of liability help account for the fact that in recent decades European auto manufacturers have been the first to introduce, or to have researched more aggressively, a number of new safety-enhancing technologies relating both to crashworthy structures and occupant protection (primarily seat belts but more recently side-impact airbags). In addition, McKay provides evidence that liability concerns played a role in the opposition by U.S. auto companies to regulatory efforts to mandate airbags.

Finally, Kip Viscusi and Michael Moore supplemented the anecdotal and historical evidence with a cross-sectional statistical analysis of the innovation effects of recent liability trends. They measured liability by annual insurance losses experienced in various sectors of the economy during the 1980–84 period, or the years in which the "liability crisis" of the 1980's first manifested itself. Innovation was proxied both by R&D costs and a measure of new products introduced. In brief, they found that for industries with relatively low liability costs, the liability system appeared, if anything, to encourage innovation—a result consistent with the theoretical arguments that liability will induce efforts by private actors to take greater precautions. But for industries such as private aircraft, where liability costs rose substantially during the 1980's and represented significant shares of total costs, liability appears to have dampened innovation.

IV. Policy Implementation

At the risk of oversimplification, the evidence compiled by the Brookings project at the very least raises significant questions about the magnitude of the "safety benefits" of the current liability system, while also providing circumstantial evidence that in certain industries its deleterious effects on innovation may be substantial. In a perfect world, therefore, one would like to scale

back its impact in just those sectors of the economy where the costs seem to outweigh the benefits.

But the U.S. legal system does not permit such precision; legal rules adopted for one industry tend, with few exceptions, to be applicable to all. Given the continued shortage of decisive empirical information, what should be done?

First, liability (as does regulation) attempts to induce safety-enhancing behavior by punishing or curtailing the activities of firms and individuals. Since this *negative* approach has at best ambiguous net benefits, it would be desirable for tort law to develop *positive* incentives for safety. Along these lines, judges and legislators might consider allowing manufacturers who demonstrate that they promptly correct design defects, or have active programs under way, to experiment with safety innovations at least some exemption from liability claims. Similar incentives should be applied for those who comply with applicable regulatory standards in a timely fashion.

Second, many of our authors pointed to the uncertainty generated by the liability system as one of its major vices. Although much of the uncertainty is due to nondoctrinal matters (differences in judges and juries, for example), one doctrinal change yet to be adopted in many jurisdictions is a "statute of repose" that would effectively immunize products from liability if they have been on the market for a lengthy period (say 10 years) without a successful tort challenge.

Third, rules of evidence in tort cases should make it possible for judges and juries to give more consideration to the broader *social* benefits of products and services at issue in tort cases, especially those at the technological frontier. The cutting-edge new drug or surgical procedure may offer many gravely ill patients a chance of recovery that cannot be found elsewhere; society at large may benefit from such technological advances even if a specific plaintiff may not. Legal rules, jury instructions, and evidential standards should thus be drafted to give more weight to the actual and potential benefits of innovations so as not to deter them with negative liability judgments.

This last point highlights the fact much of the impact of the current U.S. tort system is felt not through specific doctrines but through a variety of ancillary factors that have provided strong incentives to injured parties to use our tort law as a device for obtaining compensation far more intensively than victims in other industrialized countries. As Gary Schwartz points out in his contribution, there are surprisingly few doctrinal differences in the tort law of the United States, Europe, and Japan: all countries use a negligence standard for medical practice and "strict liability" for product design defects (which, in practice, are tried under a negligence-type standard that involves the weighing of risks against benefits). Instead, what makes the United States so different is the far greater numbers of tort actions, and the damage awards and settlements the successful suits produce. Several factors, mostly outside the narrow field of tort law, are responsible: the use of juries to decide cases, but far less frequently (or not at all) abroad; the American system of contingent attorneys' fees (present in Japan, but not in England where in fact losing parties pay the winner's lawyers fees); the more liberal rules allowing plaintiffs to recover damages for pain and suffering (the one tort-specific doctrine that may be important); and more liberal discovery rules in the United States for all types of litigation. The importance of these nontort factors, which are difficult if not impossible to change, provides a strong caution to those who believe that just by modifying certain tort doctrines, the net benefits of the overall system can be significantly enlarged.

REFERENCES

- Eads, George and Reuter, Peter, *Designing Safer Products: Corporate Responses to Product Liability Law and Regulation*, Rand Corporation, Institute for Civil Justice, 1983.
- Huber, Peter and Litan, Robert E., *Corporate Cold Feet? The Impact of Liability Law on Safety and Innovation*, Washington: Brookings Institution, forthcoming 1991.
- McGuire, E. Patrick, *The Impact of Product*

- Liability*, New York: The Conference Board, 1988.
- Peterson, Mark**, *Civil Juries in the 1980s: Trends in Jury Trial and Verdicts in California and Cook County, Illinois*, Rand Corporation, Institute for Civil Justice, 1987.
- Priest, George L.**, "Products Liability Law and the Accident Rate," in R. E. Litan and C. Winston, eds., *Liability: Perspectives and Policy*, Washington: Brookings Institution, 1989.
- Tillinghast**, *Tort Cost Trends: An International Perspective*, Simsbury, CT, 1989.
- Weber, Nathan**, *Product Liability: The Corporate Response*, New York: The Conference Board, 1989.
- U.S. Department of Justice**, *Report of the Tort Policy Working Group on the Causes, Extent and Policy Implications of the Current Crisis in Insurance Availability and Affordability*, Washington: USGPO, 1986.

Mispriced Equity: Regulated Rates for Auto Insurance in Massachusetts

By B. GLENN BLACKMON, JR. AND RICHARD ZECKHAUSER*

From the Santa Monica Freeway to the New Jersey Turnpike, drivers are unhappy about the cost of automobile insurance and are asking government to do something about it. California voters approved Proposition 103 in 1988; it requires that all rates be approved by the state insurance commissioner, attempts to reduce rates by 20 percent, and dramatically limits the criteria that can be used to rate drivers for premium purposes. New Jersey enacted an insurance reform law that seeks to charge insurers for a deficit-burdened state underwriting pool and prohibits the use of age, sex, and marital status in rating drivers for premiums. Other states enacting or considering significant rate rollbacks or reform since 1988 include Arizona, Florida, Michigan, Nevada, and Pennsylvania.

This article describes the current consequences of similar policies adopted in Massachusetts more than a decade ago. The experience suggests that recent moves by other states in the same direction will ultimately prove quite expensive as the proportion of high-cost drivers increases and as insurers lose the incentive to write policies and control costs. The trend away from insurance premiums based on expected cost also reduces incentive effects for drivers, since insurance premiums provide a link between tort judgments and consumer decisions.

I. Insurance Regulation in Massachusetts

The insurance commissioner in Massachusetts specifies a rating system for classifying drivers and sets a single schedule of

rates that apply to any insurer in the state. The state eliminated sex-based rate differences for automobile insurance in 1977, when the state insurance commissioner concluded that "sex classifications in automobile insurance represent unfair discrimination. Rates containing a distinction based on gender are both unjust and violative of public policy" (James Stone, 1978, p. 179). At the same time, rate differences based on age were also eliminated and replaced with rating classes based on the number of years as a licensed driver. The state classification system allows rates to vary by territory (i.e., the place of residence of the driver), class (i.e., whether an inexperienced driver is the principal or an occasional driver of the car, whether any inexperienced driver has had driver training, and whether the car is used for business), and the age and type of car.

The state has "tempered" (limited the variation of) premiums across territories and classes. Across classes, the expected cost of insuring a driver varies by a factor of 4.4, but premiums vary by a factor of 3. Across territories, costs vary by 2.7 and premiums by 2. When class and territory are combined, the effect of tempering is even greater. The premium for drivers in the highest class-territory cell is 4.50 times the premium of the driver in the lowest cell, yet the cost of insuring the high-cell driver is 10.6 times the cost of insuring the low-cell driver (Automobile Insurers Bureau, 1990, exhibit 5).

In addition to the transfers among drivers, insurance regulators have also apparently tried to effect a transfer from insurers to drivers as a group. The insurance industry contends that overall rates are consistently below the level necessary to provide a profit. This claim would be expected from the industry as strategic behavior in the rate-setting process, and it would have little credibility if insurers were continuing to do

*John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138. We thank the Bradley Foundation for research support, and John Pratt.

business in the state. Yet by 1989, eight insurance companies, writing 25 percent of automobile coverage in the state, had abandoned business in Massachusetts or were actively attempting to do so, despite high exit fees imposed by the insurance commissioner. These exit fees include both cash payments and prohibitions on writing other lines of insurance in the state.

While an insurance company is allowed no flexibility in the rates it charges, it can choose whether to insure a particular risk. Many companies choose not to insure many drivers. If an insurer declines to insure an exposure, that driver is ceded to a residual market pool, Commonwealth Auto Reinsurers (CAR). Being ceded to the pool has no effect on the driver's premium, and as a result the residual market pool pays claims well in excess of premiums. In 1988, 63 percent of drivers were insured by CAR, and the CAR deficit for private passenger automobile insurance was \$519 million, or \$239 per ceded risk (Commonwealth Automobile Reinsurers, 1988). This residual market deficit is financed by surcharging premiums for drivers in the voluntary market.

II. Subsidies in Massachusetts Auto Insurance Rates

The subsidies contained in Massachusetts auto insurance rates are large and multidimensional. Subsidies flow from drivers in rural areas to those in urban areas, from women to men, from the middle-aged to the young and the elderly, from experienced drivers to inexperienced drivers, and from drivers in the voluntary market to those in the high-risk, involuntary market.

How large is the subsidy? A minimum value would be the size of the residual market deficit (\$519 million), but the true value is surely larger since some ceded drivers are actually paying more than their cost. About \$191 million results from the tempering of rates across class/territory cells. The remainder, at least \$328 million, can be attributed to the state rating system that requires insurers to charge a single premium

TABLE 1—AVERAGE SUBSIDY AND PRICE OF THOSE PAYING AND RECEIVING SUBSIDIES

	Average Cost	Insured Vehicles (000)	Average Premium	Subsidy
<i>s</i>	\$1,312	820	\$1,079	\$233
<i>c</i>	\$928	1,251	\$665	\$262
<i>n</i>	\$323	1,216	\$750	(\$427)

Source: Authors' calculations from data in Automobile Insurers Bureau (1990).

Note: *s* = subsidized cell; *c* = ceded to residual market; *n* = not ceded or subsidized.

to a heterogeneous group of drivers. A cell that, on average, is charged premiums in excess of cost will nonetheless contain some drivers whose costs exceed the premium. Many of these drivers can be readily identified by an insurer and are ceded to the residual pool. For example, the average experienced driver 25 or older pays a subsidy of about \$60 per year, yet 60 percent of these drivers are ceded to the residual market. Recognizing this, we divide drivers into three groups in Table 1: those in a cell that receives a subsidy through tempering (*s*); those in a subsidy-paying cell but who receive a subsidy from ceding (*c*); and those neither tempered nor ceded (*n*). The first two groups receive a subsidy, and the third group pays a subsidy. We assume that all drivers in a subsidy-receiving cell are ceded and that the proportion of ceded drivers is constant across subsidy-paying cells.

These subsidies generate allocative inefficiency: those who pay the subsidies restrict their consumption of automobile insurance, by not driving or by driving without insurance. Those receiving the subsidies increase their consumption. A deadweight loss results as some consumers are deterred from driving even though they would pay the cost and others drive when they would not if prices reflected costs. To assess the deadweight loss from the subsidies, we first estimated a simple demand function for insurance using data from Massachusetts towns in 1988. The demand for insured vehicles per household was estimated as a log-linear

(constant-elasticity) function of income, price, and household density.¹

Using this demand function, we calculated deadweight loss from the subsidies at \$217 million annually, or 42 percent of the total subsidy. (This 42 percent figure seems high until we recognize that if the cross-subsidy scheme must break even, the losses of both those undercharged and overcharged must be counted.)² The deadweight loss is substantial, but it alone does not tell us whether the policy of subsidies is good or bad. There are at least two possible rationales for subsidies—risk spreading and egalitarianism.

Subsidies as a risk-spreading mechanism. The state's rationale for subsidizing high-cost drivers at the expense of low-cost drivers (as explained by the commissioner) in essence is to spread the risk that one may be a bad driver (Stone). If drivers were risk averse and did not yet know whether they are high- or low-cost drivers, they could increase their expected utility by agreeing to

pay more than cost if they are low-cost and less than cost otherwise. How risk averse would Massachusetts drivers have to be for the observed pattern of tempering to be superior to cost-based prices? To answer this question, we first computed the gain from driving, g , as the per capita consumers' surplus up to the \$5,000 available at the bottom of the demand curve. We then posited household utility, U , as a function of money income, y , plus the gain g derived from operating an insured vehicle. For each cell of individuals, i , let $x_i = y_i + g_i$. We express utility as a constant proportional risk-aversion function, $U_i = u(x_i) = x_i^b$, where b is a risk-aversion parameter to be estimated. We measure the effect of the subsidies by their effect on the expected utility of a randomly chosen individual, one who knows only the proportions of individuals who will be subsidized and by how much but not his particular group. The relevant individuals are those who would own a car under either the subsidized rates or cost-based rates. (There would be 1.993 million drivers in group n if rates were based on cost.)

For any given set of prices, the expected utility $E(U)$ is equal to the utility of an individual in each group, weighted by the proportion of individuals (potential drivers) in that group. Each of the three groups in Table 1 is divided into finer units based on class and territory, yielding a total 1,092 cells. Thus

$$E(U) = \sum_{i=1}^{1,092} [d_i * (30,000 + g_i)^b],$$

where d_i is the proportion of individuals in cell i and all drivers are assumed to have household income of \$30,000. We find the highest value for b at which $E(U)$ under the current subsidy regime is equal to $E(U)$ under a regime in which each group was charged a price equal to its expected cost. For the subsidized outcome to offer at least as high an expected utility, b must be less than or equal to -7.2 . At this level of risk aversion, an individual's marginal utility of

¹Our illustrative estimation applies ordinary least squares to data from 294 towns. Our right-hand side variables were median household income in 1979 (the latest year of town-level data), average price of a standard package of insurance coverage, and households per square mile as the density measure. The demand for vehicles should decrease with population density because substitute transportation becomes more readily available. Thus, we expect the coefficients on income to be greater than zero, and on price and density to be less than zero. Our estimated coefficients, with standard errors in parentheses, were income .477 (.044), price $-.569$ (.119), and density $-.044$ (.011). Our R^2 was .593. All coefficients are significantly different from zero at the 99 percent confidence level (see our 1990 paper). We did not have a sufficiently rich data set to control for exogenous variables that could influence the driving decision. In contrast to a traditional market, supply conditions were favorable for our estimation efforts, since all prices were set by regulators, and the supply curve for each cell was horizontal (any driver had to be accommodated at the established rates).

²Our measure of deadweight loss does not include the inefficiency from pricing at average cost instead of marginal cost. Using data on insurance costs, we estimate that marginal cost exceeds average cost by 40–60 percent in urban areas of Massachusetts. The additional deadweight loss is \$81 million (see our 1990 paper).

income drops by an implausible factor of 295 as his income increases from \$15,000 to \$30,000.

Subsidies as an income transfer mechanism. A second rationale for subsidies would be to transfer income from consumers with high income to those with low income, again as a way of raising the expected utility of a randomly selected consumer (i.e., egalitarianism in the spirit of an optimal income tax trading off incentive losses against risk-spreading gains). Tempering a commodity price, such as an auto insurance premium, can accomplish this purpose only if high-cost consumers tend to have low incomes. The relationship between subsidies and income is decidedly mixed in the case of automobile insurance. The subsidy of Boston and other cities tends to flow from high-income towns to low-income towns. Yet women subsidize men even though women's income is much lower. In many cases the group paying or receiving a subsidy is simply too diverse for us to estimate its income.

Measuring only variations across towns, we estimate that the average household income of group *s*, those in a cell receiving a tempering subsidy, is \$26,500, measured in 1988 dollars. Since the drivers in group *c* and group *n* are in the same cells, their average income is the same, \$30,418. Given these assumptions about income, an egalitarian motive justifies the subsidies if $b < -2.7$. At this level of risk aversion/egalitarianism, the marginal utility of income at \$15,000 is 13 times that at \$30,000. (For this utility function, taking account of variability in incomes within cells would increase the relative attractiveness of the subsidy scheme.)

We have not attempted to measure the egalitarian sentiment of Massachusetts citizens, but a look at the state's social programs and tax system suggests that it is not nearly this strong. Moreover, this redistributive bucket is extraordinarily leaky; other instruments, such as taxes and transfers, could effect the transfer with significantly smaller losses.

"Price equity" as a goal. Neither risk aversion nor egalitarianism, if their implicit

tradeoff rates are examined, yields a satisfactory justification for the subsidies in Massachusetts auto insurance rates. We believe that the subsidies may be better explained (though not justified) by a desire for price equity, the idea that differences in price for goods that are nominally similar is in itself a bad thing. Price equity is a sensible goal, and a natural accompaniment of competitive markets, when the cost of the product does not depend on who consumes it. Variation in prices across a market is usually welfare reducing, but not for insurance. The expected cost of insurance varies in predictable fashion for large subsets of consumers, though not for any particular consumer.

Even a well-informed consumer, however, would find it difficult to judge the absolute or relative cost of his coverage and would probably underestimate variation in cost across insureds. With little idea about cost, consumers judge their rate relative to others—both rates charged other consumers and the rates they paid in the past. If consumers assess the fairness of rates on a relative scale, a politically responsive regulator would set rates on the same basis.

The recent insurance reforms in California and New Jersey significantly temper rates, promoting the notion that significant variation in rates across consumers is unfair. The demand for price equity is reflected in prices for electricity and natural gas (typically the same for rural and urban customers of the same company despite cost differentials), tuitions for college students (invariant across fields with disparate costs of facilities and faculty), and the invariant cost of sending a letter.

III. Regulation and the Residual Market

A heavily subsidized market has important and negative effects on the incentives of insurers. These effects are most apparent in the operation of the residual or involuntary market, which was intended to be a last-resort insurance source for high-risk drivers. It now insures more than two-thirds of Massachusetts drivers, its size the inevitable consequence of that state's strategy

of limiting absolute rates and tempering relative rates.

Every state provides some mechanism to insure drivers who cannot obtain insurance at standard rates. The most common mechanism, used in about 40 states, is the assigned-risk plan: drivers who cannot obtain insurance are assigned to an insurer, who is responsible for all expenses and losses of that policy. Massachusetts uses a reinsurance mechanism. Drivers are not assigned to a specific insurer; rather their losses are allocated among insurers. Massachusetts is unique in requiring that residual and voluntary market drivers in the same rating category pay the same rate. Tempering and heterogeneous classifications both increase the ceding of risks, and Massachusetts now has more drivers in the involuntary market (70 percent of all drivers in 1989) than any other state (Timothy Gailey, 1989, p. 5). Since 1977, when the state adopted its policy of tempering rates and prohibiting sex- and age-based rates, the total number of insured drivers has increased by 35 percent. However, the number of drivers voluntarily covered by insurance companies has decreased by 35 percent.

The high proportion of drivers in the residual pool undermines the incentive of insurers to minimize cost; for instance, by limiting fraudulent claims and excessive payments for repairs. Most claims are paid from the residual pool, where all insurers share the loss. Thus an individual insurer will bear all of the cost and almost none of the benefit from greater efforts to control the cost of these claims.

IV. The Implication for Torts

The systematic mispricing of automobile insurance in Massachusetts, and the trend toward the Massachusetts method in other

states, undermines the rationale for preserving some element of a tort liability system. When insurance is priced at cost, the premium reflects the tort claims that can be expected if the consumer engages in the insured activity, and the consumer will make an efficient decision about whether to operate a motor vehicle, though his incentive to drive safely will be insufficient.

V. Conclusion

Massachusetts suffers from significant deadweight efficiency losses, hence high prices, because regulated rates for auto insurance deviate substantially from cost. Taking an expected utility approach, for reasonable parameter values neither risk-spreading nor egalitarian concerns justify this cross-subsidy scheme.

REFERENCES

- Blackmon, B. Glenn, Jr. and Zeckhauser, Richard J., "Mispriced Equity: Automobile Insurance Pricing in Massachusetts," mimeo., John F. Kennedy School of Government, Harvard University, 1990.
- Gailey, Timothy H., *CAR Reform: Shaping the Residual Market for the 1990's*, Boston: Massachusetts Division of Insurance, 1989.
- Stone, James M., "Excerpt from the Opinion, Findings and Decision on 1978 Automobile Insurance Rates," in *Automobile Insurance Risk Classification: Equity and Accuracy*, Boston: Massachusetts Division of Insurance, 1978.
- Automobile Insurers Bureau of Massachusetts, *Subsidies in the 1990 Rates*, Actuarial Notice 90-1, Boston: 1990.
- Commonwealth Automobile Reinsurers, *Members Participation Report, 1988*, Boston: CAR, 1988.

PATH DEPENDENCE IN ECONOMICS: THE INVISIBLE HAND IN THE GRIP OF THE PAST[†]

Multiple Equilibria and Persistence in Aggregate Fluctuations

By STEVEN N. DURLAUF*

Recent developments in theoretical macroeconomics have emphasized the potential for multiple, Pareto-rankable equilibria to exist for economies where various Arrow-Debreu assumptions are violated. Authors such as Peter Diamond (1982) emphasized how incomplete markets can allow economies to become trapped in Pareto-inferior equilibria; Walter Heller (1986) obtained similar results due to imperfect competition. These different approaches share the idea that strong complementarities in behavior can lead to multiplicity. Intuitively, when technological or demand spillovers make agents sufficiently interdependent, high and low levels of activity can represent internally consistent equilibria in the absence of complete, competitive markets. Most of these models describe multiple steady states in economies rather than multiple nondegenerate time-series paths, and consequently cannot address issues of aggregate fluctuations. Further, this literature has not shown how economies can shift across equilibria, inducing periods of boom and depression.

An independent literature has argued that aggregate fluctuations are strongly persistent. Researchers have concluded from a variety of perspectives that aggregate output in advanced industrialized economies con-

tains a unit root. Despite controversy over the exact magnitude of the permanent component, the effects of current events on real activity apparently persist over long horizons.

The purpose of the current paper is to link the new multiplicity results in macroeconomic theory with the evidence on output persistence. I do this by modeling coordination problems in an explicitly stochastic framework. As developed in my earlier papers (1990; 1991), the microeconomic specification of the economy is expressed as a set of conditional probability measures describing how individual agents behave given the economy's history. An aggregate equilibrium exists when one can find a joint probability measure over all agents which is consistent with these conditional measures; multiplicity occurs when several such measures exist. This approach permits one to directly describe the time-series properties of aggregate fluctuations along different equilibrium paths.

Specifically, I examine the capital accumulation problems of a set of infinitely lived industries. I deviate from standard analyses in two respects. First, each industry faces a nonconvex production technology. Second, industries experience technological complementarities as past high production decisions by each industry increase the current productivity of several industries through dynamic learning by doing or other effects. Industries do not coordinate production decisions because of incomplete markets. By describing how output levels and productivity evolve as industries interact over time, the model characterizes the impact of complementarities and incomplete markets on the structure of aggregate fluctuations.

[†]*Discussants:* W. Brian Arthur, Stanford University; Paul A. David, Stanford University; Paul Romer, University of California-Berkeley; Robert M. Solow, MIT.

*Department of Economics, Stanford University, Stanford, CA 94305. I thank Charles Bean, Suzanne Cooper, Paul David, Avner Greif, Robert Solow, Doug Steigerwald, and Jeroen Swinkels for helpful comments.

I. A Model of Interacting Industries

Consider a countable infinity of industries indexed by i .¹ Each industry consists of many small, identical firms. All firms produce a homogeneous good; industries are distinguished by distinct production functions rather than distinct outputs. The homogeneous final good may be consumed by the owners of the firms, or converted to a capital good that fully depreciates after one period. Industry i 's behavior is proportional to the behavior of a representative firm that chooses a capital stock sequence $\{K_{i,t}\}$ to maximize the present discounted value of profits $\Pi_{i,t}$

$$(1) \quad \Pi_{i,t} = E \left(\sum_{j=0}^{\infty} \beta^j (Y_{i,t+j} - K_{i,t+j}) | \mathfrak{F}_t \right),$$

where $Y_{i,t}$ equals the output of the i th industry's representative firm at t ; \mathfrak{F}_t equals all available information at t . Initial endowments $Y_{i,0}$ provide starting capital.

Aggregate behavior is determined by the interactions of many heterogeneous industries employing nonconvex technologies. Production occurs with a one-period lag; firms at $t-1$ employ one of two production techniques and a level of capital to determine output at t . Only one technique may be used at a time. Russell Cooper (1987) and Kevin Murphy et al. (1989) exploit similar technologies to analyze multiple equilibria; Paul Milgrom and John Roberts (1990) discuss how this type of nonconvexity can arise as firms internally coordinate many complementary activities. The technique-specific production functions produce $Y_{1,i,t}$ and $Y_{2,i,t}$ through

$$(2) \quad Y_{1,i,t} = f_1(K_{i,t-1} - F_i, \zeta_{i,t-1}, \xi_{t-1})$$

$$Y_{2,i,t} = f_2(K_{i,t-1}, \eta_{i,t-1}, \xi_{t-1}).$$

$\zeta_{i,t}$ and $\eta_{i,t}$ are industry-specific productivity shocks; ξ_t is an aggregate productivity

shock and F_i is a fixed overhead capital cost; $\zeta_{i,t-1}$, $\eta_{i,t-1}$, and ξ_{t-1} are elements of \mathfrak{F}_{t-1} . Recalling that firms within an industry are identical, let us define $\omega_{i,t}$ which equals 1 if technique 1 is used by industry i at t , 0 otherwise and $\omega_t = \{\dots \omega_{i-1,t}, \omega_{i,t}, \omega_{i+1,t}, \dots\}$ which equals the joint set of techniques employed at t .

I make the following assumptions. First, each technique fulfills standard curvature conditions. Further, I associate technique 1 with high production. Specifically, net capital $NK_{i,t}$, that equals $K_{i,t} - F_i$ for technique 1 and $K_{i,t}$ for technique 2, has a strictly higher marginal (and by implication total) product when used with technique 1 than technique 2. A firm chooses technique 1 if it is willing to pay fixed capital costs in exchange for higher output.

ASSUMPTION 1: Restrictions on technique-specific production functions.

For all realizations of $\zeta_{i,t}$, $\eta_{i,t}$, and NK ,

- A. $f_1(0, \zeta_{i,t}, \xi_t) = f_2(0, \eta_{i,t}, \xi_t) = 0$.
- B. $\frac{\partial f_1(0, \zeta_{i,t}, \xi_t)}{\partial NK} = \frac{\partial f_2(0, \eta_{i,t}, \xi_t)}{\partial NK} = \infty$;
 $\frac{\partial f_1(\infty, \zeta_{i,t}, \xi_t)}{\partial NK} = \frac{\partial f_2(\infty, \eta_{i,t}, \xi_t)}{\partial NK} = 0$.
- C. $\frac{\partial f_1(NK, \zeta_{i,t}, \xi_t)}{\partial NK} > \frac{\partial f_2(NK, \eta_{i,t}, \xi_t)}{\partial NK}$.

Both techniques are assumed to exhibit technological complementarities, as the history of realized activity determines the parameters of the production function at t . Paul Romer's (1986) model of social increasing returns shares this feature. My complementarities differ from Romer's in two respects. First, all complementarities are local as the production function of each firm is affected by the production decisions of a finite number of industries. The index i orders industries by similarity in technology; spillovers occur only between similar technologies. Paul David (1988) describes the historical importance of local complemen-

¹My 1990 paper derives a general equilibrium version of this economy.

tarities in the evolution of technical innovations. Second, my complementarities are explicitly dynamic. Past production decisions affect current productivity, which captures the idea of learning by doing.

Specifically, I model the complementarities through the dependence of the productivity shocks $\zeta_{i,t}$ and $\eta_{i,t}$ on the history of technique choices (see my 1990 paper for a justification). Complementarities are assumed to be the only source of dependence across shocks. $\text{Prob}(x|y)$ denotes the conditional probability measure of x given information y ; $x(y)$ denotes the random variable associated with this measure. $\Delta_{k,l} = \{i - k \dots i \dots i + l\}$ indexes the industries that affect industry i 's productivity.

ASSUMPTION 2: Conditional probability structure of productivity shocks.

A. $\text{Prob}(\zeta_{i,t}|\mathfrak{F}_{t-1})$

$$= \text{Prob}(\zeta_{i,t}|\omega_{j,t-1} \forall j \in \Delta_{k,l}).$$

B. $\text{Prob}(\eta_{i,t}|\mathfrak{F}_{t-1})$

$$= \text{Prob}(\eta_{i,t}|\omega_{j,t-1} \forall j \in \Delta_{k,l}).$$

C. The random pairs $(\zeta_{i,t} - \zeta_{i,t}(\mathfrak{F}_{t-1}), \eta_{i,t} - \eta_{i,t}(\mathfrak{F}_{t-1}))$ are mutually independent of each other and of $\xi_i - \xi_i(\mathfrak{F}_{t-1}) \forall i$.

No markets exist whereby individual firms can coordinate to exploit complementarities. Consequently, no industry may be compensated for choosing technique 1 in order to expand the production sets of other industries; nor, given my conceptualization of industries as aggregates of many small producers, can firms within an industry strategically choose a technique in order to induce higher future productivity through complementarities. Market incompleteness combines with the production nonconvexity to fundamentally affect aggregate dynamics.

II. Local Complementarities and Multiple Equilibria

I initially analyze the economy without aggregate shocks, by setting $\xi_i = 0 \forall i$. From my assumptions, one may show that equilib-

rium industry technique choices obey conditional probabilities of the form

$$(3) \quad \text{Prob}(\omega_{i,t}|\mathfrak{F}_{t-1}) \\ = \text{Prob}(\omega_{i,t}|\omega_{j,t-1} \forall j \in \Delta_{k,l}).$$

Once technique choices are determined, one can solve for the optimal levels of capital and output for each firm. In fact, a sufficient condition for the existence of equilibrium capital and output sequences for all firms is the existence of a joint probability measure over all technique choices which is consistent with the conditional measures (3). My 1990 paper verifies that such a joint measure exists for any initial conditions ω_0 .

Let us now restrict the conditional probabilities in order to discuss multiplicity and dynamics. Past choices of technique 1 are assumed to improve the current relative productivity of the technique. As a result, technique 1 choices will propagate over time. Further, it is assumed that $\omega_i = 1$ is a steady state, which means that when all productivity spillovers are active, the effects are so strong that high production is always optimal.

ASSUMPTION 3: Impact of past technique choices on current technique probabilities.²

Let ω and ω' denote two realizations of ω_{i-1} . If $\omega_j \geq \omega'_j \forall j \in \Delta_{k,l}$, then

A. $\text{Prob}(\omega_{i,t} = 1|\omega_{j,t-1} = \omega_j \forall j \in \Delta_{k,l})$

$$\geq \text{Prob}(\omega_{i,t} = 1|\omega_{j,t-1} = \omega'_j \forall j \in \Delta_{k,l}).$$

B. $\text{Prob}(\omega_{i,t} = 1|\omega_{j,t-1} = 1 \forall j \in \Delta_{k,l}) = 1.$

Whenever some industry chooses $\omega_{i,t} = 0$, a positive productivity feedback is lost. Different configurations of choices at $t-1$ determine different production sets and conditional technique choice probabilities for

²This assumption can be reformulated in terms of restrictions on the technique-specific production functions.

each industry. I bound the technique choice probabilities from below and above by $\Theta_{k,l}^{\min}$ and $\Theta_{k,l}^{\max}$, respectively.

$$(4) \quad \Theta_{k,l}^{\min} \leq \text{Prob}(\omega_{i,t} = 1 | \omega_{j,t-1} = 0$$

$$\text{for some } j \in \Delta_{k,l}) \leq \Theta_{k,l}^{\max}.$$

Since $\omega_i = 1$ is an equilibrium, multiple equilibria exist if for some initial conditions, $\omega_i = 1$ fails to emerge as t grows. Notice that even if $\omega_0 = 0$, favorable productivity shocks will periodically induce industries to produce using technique 1. The choice of technique 1 by one industry, through the complementarities, increases the probability that the technique is subsequently chosen in several industries. With strong spillovers, these effects may build up, allowing $\omega_i = 1$ to emerge from any initial conditions. The model therefore allows us to analyze the stability of a high aggregate output equilibrium from arbitrary initial conditions.

In fact, the limiting behavior of the economy is determined by the bounds $\Theta_{k,l}^{\min}$ and $\Theta_{k,l}^{\max}$. If the probability of high production by an industry is sufficiently large for all production histories, then the spillover effects induced by spontaneous technique 1 choices cause the economy to iterate towards high production. Alternatively, if technique 1 probabilities are too low in the absence of active spillovers, spontaneous technique 1 choices will not generate sufficient momentum to achieve the $\omega_i = 1$ equilibrium. $\Theta_{k,l}^{\min}$ and $\Theta_{k,l}^{\max}$ bound the degree of complementarity in the economy. Large values of $\Theta_{k,l}^{\min}$ imply complementarities are weak as technique 1 is chosen relatively frequently regardless of the past. Conversely, small values of $\Theta_{k,l}^{\max}$ imply strong complementarities; the probability of current high production is very sensitive to past technique choices. Theorem 1 (proven in my 1990 paper) shows how long-run industry behavior is jointly determined by initial conditions and conditional technique probabilities.

THEOREM 1: *Conditions for uniqueness vs. multiplicity of long-run equilibrium.*

For every nonnull index set $\Delta_{k,l}$, there exist numbers $0 < \underline{\Theta}_{\Delta_{k,l}} < \bar{\Theta}_{\Delta_{k,l}} < 1$ such that

A. If $\Theta_{k,l}^{\max} \leq \underline{\Theta}_{\Delta_{k,l}}$, then

$$\lim_{t \rightarrow \infty} \text{Prob}(\omega_{i,t} = 1 | \omega_0 = 0) < 1.$$

If complementarities are sufficiently strong, no industry converges to the high production technique almost surely from economywide low production technique initial conditions.

B. If $\Theta_{k,l}^{\min} \geq \bar{\Theta}_{\Delta_{k,l}}$, then

$$\lim_{t \rightarrow \infty} \text{Prob}(\omega_{i,t} = 1 | \omega_0 = 0) = 1.$$

If complementarities are sufficiently weak, each industry converges to the high production technique almost surely from economywide low production technique initial conditions.

One can associate $\omega_i = 1$ with the equilibrium that would emerge if all firms chose their production levels cooperatively. If production through technique 1 is sufficiently large for $\omega_i = 1$ versus any other configuration, then $\omega_i = 1$ emerges as the cooperative equilibrium after one period. Consequently, incompleteness of markets lowers the mean and increases the variance of industry and aggregate output along the inefficient equilibrium path, as technique choices fluctuate over time. When industries fail to coordinate, production decisions become dependent on idiosyncratic productivity shocks. Observe that the volatility associated with the inefficient equilibrium is caused by fundamentals. Simulations in both of my earlier papers show that aggregate output can obey a wide range of AR processes, depending on $\Delta_{k,l}$.

III. Path Dependence and Aggregate Shocks

Now consider the role of the aggregate shocks ξ_t . By affecting many industries simultaneously, these shocks act in a way analogous to changing the initial conditions of the economy. Path dependence occurs as

one realization of ξ_t permanently changes the equilibrium in the absence of future offsetting shocks. I assume that sufficiently unfavorable aggregate productivity draws make technique 1 unlikely whereas sufficiently favorable draws ensure the use of the technique.

ASSUMPTION 4: *Impact of aggregate shocks on technique choice.*

There exist numbers a and b , with $\text{Prob}(\xi_t \leq a)$ and $\text{Prob}(\xi_t \geq b)$ both nonzero, such that³

$$A. \quad \text{Prob}(\omega_{i,t} = 1 | \xi_t \leq a, \omega_{j,t-1} = 1 \forall j \in \Delta_{k,i}) \\ \leq \underline{\Theta}_{\Delta_{k,i}}.$$

$$B. \quad \text{Prob}(\omega_{i,t} = 1 | \xi_t \geq b, \omega_{j,t-1} = 0 \forall j \in \Delta_{k,i}) \\ = 1.$$

When this assumption holds, aggregate shocks can have an indefinite effect on real activity. My 1991 paper verifies

THEOREM 2: *Path dependence due to aggregate shocks.*

Let $\xi_t = 0 \forall t > T$ and $\Theta_{k,i}^{\max} \leq \underline{\Theta}_{\Delta_{k,i}}$. The economy exhibits path dependence as the realization of ξ_T affects the limiting technique choice probabilities for all industries.

$$A. \quad \lim_{t \rightarrow \infty} \text{Prob}(\omega_{i,t} = 1 | \xi_T \leq a) < 1.$$

$$B. \quad \lim_{t \rightarrow \infty} \text{Prob}(\omega_{i,t} = 1 | \xi_T \geq b) = 1.$$

This result shows how fluctuations can be persistent. For example, once many sectors simultaneously decline due to an adverse aggregate shock, productivity-enhancing complementarities are lost until a subsequent favorable draw restores them. If ξ_t is ergodic, then the economy will cycle between the equilibria.

Several interpretations beyond productivity can be applied to the aggregate shocks. Interpreting ξ_t as a proxy for the financial sector, the model indicates how the breakdown of financial institutions, such as occurred during the Great Depression, can cause indefinite output loss. Alternatively, my 1990 paper shows how ξ_t can represent the cost of production inputs provided by leading sectors such as transportation or steel. In this case, the growth of leading sectors improves the relative profitability of high production, which can lead to a takeoff in growth as the economy shifts across equilibria.

REFERENCES

- Cooper, Russell, "Dynamic Behavior of Imperfectly Competitive Economies with Multiple Equilibria," NBER Working Paper No. 2388, 1987.
- David, Paul A., "Path-Dependence: Putting the Past in the Future of Economics," Stanford University, 1988.
- Diamond, Peter A., "Aggregate Demand In Search Equilibrium," *Journal of Political Economy*, October 1982, 90, 881-94.
- Durlauf, Steven N., "Nonergodic Economic Growth," Stanford University, 1990.
- , "Path Dependence in Aggregate Output," Stanford University, 1991.
- Heller, Walter P., "Coordination Failure Under Complete Markets with Applications to Effective Demand," in his et al., *Essays in Honor of Kenneth J. Arrow*, Vol. II, Cambridge: Cambridge University Press, 1986.
- Milgrom, Paul and Roberts, John, "The Economics of Modern Manufacturing: Technology, Strategy and Organization," *American Economic Review*, June 1990, 80, 511-28.
- Murphy, Kevin, Shleifer, Andrei and Vishny, Robert, "Industrialization and the Big Push," *Journal of Political Economy*, October 1989, 97, 1003-26.
- Romer, Paul M., "Increasing Returns and Long Run Growth," *Journal of Political Economy*, October 1986, 94, 1002-37.

³The specified bound in statement A below is defined in Theorem 1.

Identifying the Hand of Past: Distinguishing State Dependence from Heterogeneity

By JAMES J. HECKMAN*

A basic problem in social science is to ascertain the importance of initial endowments on subsequent outcomes of a dynamic process. Interest in this topic centers on two distinct issues: (a) Do initial endowments have a temporally persistent effect on outcomes (i.e., is there "heterogeneity")?; (b) Are the effects of initial endowments attenuated or accentuated by subsequent experiences of the process being studied or by related processes (i.e., is there "state dependence")?

These two questions show up in a variety of contexts. 1) The importance of family background on a person's subsequent education and earnings is a hotly debated topic. Do market or nonmarket mechanisms reinforce or diminish initial endowments? 2) The incidence of criminal activity is concentrated among a small population of repeated offenders. Are certain persons "prone" to criminality, or does crime breed crime? 3) Are persons who experience unemployment more likely to experience future unemployment due to loss of work experience or market stigma? 4) Does early entry into an industry confer an advantage of incumbency, or does it merely proxy the basic managerial and innovative ability of early entrants?

Many formal models of state-dependent processes or processes with persistent effects of initial conditions have been formulated. Invariance of steady states to initial endowments was viewed as a desirable feature of an economic model in the 1960's and 1970's. Path-dependent synergism and nonergodicity are fashionable features of models today. Is the choice between models

with long-run dependence on initial conditions and those without solely a matter of intellectual esthetics? Can data discriminate between these two classes of models? How many economically extraneous statistical "regularity conditions" have to be imposed in order to distinguish between these models? How many and what kind of *maintained* assumptions are required in order to let the data speak on these important questions?

In this paper I focus on one nonergodic model that receives much attention in the literature—the model of "duration dependence" and "heterogeneity" discussed by Herbert Silcock (1954) and also examined by Tony Lancaster (1979), Chris Elbers and Geert Ridder (1982), myself and Burton Singer (1984), and myself and Bo Honoré (1989).

I. The Problem of Separating Heterogeneity from Duration Dependence

The length of time in a state, L , and its determinants are the objects of interest. For example, the length of time that a new entrant firm survives in an industry, the length of an unemployment spell or a job spell or a strike, the length of time a person spends in school, or the time to a first birth are all objects of interest in current research. For simplicity, assume a continuous time model.

A set of explanatory variables is thought to explain L , but in addition there is some "randomness" or unpredictability given the explanatory variables. Outcomes depend on a larger set of variables than the economic analyst has at his or her disposal. We write (X, Θ) where X is a vector of observed explanatory variables and Θ is a scalar unobserved variable known to the agent but not the econometrician. The assumption that Θ is scalar is made for analytical con-

*University of Chicago, Chicago, IL 60637. This work was supported by NSF-87-10145 and NIH grant HD-19226. I have benefited from discussions with Ricardo Barros and Bo Honoré.

venience. The assumption of "randomness" is that the distribution of L is nondegenerate given (X, Θ) , that is,

$$(1) \quad \Pr(L \leq l|x, \theta) = F(l|x, \theta).$$

If (X, Θ) fully explain durations, the probability on the left would be zero or one. The distinction between Θ and the unobservable giving rise to the nondegenerate conditional distribution of L is arbitrary. Distribution (1) is of economic interest if agents make decisions based on Θ but not on the random components that make the distribution nondegenerate given (X, Θ) . Then (1) can properly be called a structural distribution and is an object of economic interest.

In duration analysis it is convenient to work with the survivor function: $S(l|x, \theta) = \Pr(L > l|x, \theta)$, the conditional probability that a spell exceeds length l . To avoid technicalities, assume that L is continuously distributed so the density of L is well defined. The structural conditional hazard rate or exit rate from the spell is denoted $h(l|x, \theta)$. If h is increasing in l , positive duration dependence is said to be present for those values of l . If h is decreasing, negative duration dependence is said to be present. In either case, state dependence (or one form of the hand of the past) is present because the length of time spent in a state determines the conditional exit rate from the state.

For time-invariant (X, Θ) or for time-varying (X, Θ) satisfying the conditions given in my article with James Walker (1990a),

$$(2) \quad S(l|x, \theta) = \exp - \int_0^l h(u|x(u), \theta(u)) du.$$

The survivor is the inverse of the exponentiated integrated hazard.¹ For analytical convenience I assume that $\Theta(u) = \Theta$ is constant for all u and is a scalar random variable. Its distribution function is $G(\theta)$. The variable Θ is assumed statistically independent of X , and can thus be interpreted

as a time-invariant endowment or propensity.

The problem of distinguishing state dependence from heterogeneity (or initial endowments) can now be stated precisely. The data determine the conditional distribution of L given X which is the *mean* conditional survivor function:

$$(3) \quad S(l|x) = \int_{\theta} S(l|x, \theta) dG(\theta) \\ = \int_{\theta} \exp \left(- \int_0^l h(u|x(u), \theta) du \right) dG(\theta)$$

where θ is the support of Θ . Associated with $S(l|x)$ is the empirical hazard

$$(4) \quad h(l|x).$$

A standard result in the literature establishes that (4) is biased toward negative duration dependence relative to $h(l|x, \theta)$.² The more mobility-prone agents exit spells more rapidly producing a distribution of Θ conditional on L and X that is weighted toward less mobile values.

The issue is whether or not it is possible to identify $G(\theta)$ and $h(u|x(u), \theta)$ from $S(l|x)$. Without further structure imposed, it is not. There is no way to separate out the contributions of state dependence from heterogeneity in a completely general model. The data do not "speak for themselves" except in the case where $h(l|x)$ shows positive duration dependence. For this to arise, $h(l|x, \theta)$ must show positive duration dependence at least for certain intervals of the support of Θ . (See myself and Singer, 1985.)

II. Separability as an Avenue of Identifiability

One way to answer the stated question is to impose specific functional forms. My article with Singer (1984) presents identification conditions for a variety of conventional models in which $h(\cdot|\cdot)$ is specified to be-

¹For general time-varying explanatory variables, relation (2) is not true although it is commonly assumed to be true in applied work.

²See Proposition 1 in my book with Singer (1985, p. 53).

long to a parametric family and $G(\theta)$ is assumed to be a proper probability distribution but its functional form is not specified. Identification through explicit functional form assumptions about both h and G is conventional but controversial. See the examples of the consequences of misspecification discussed in my book with Singer (1985).

An alternative and somewhat more general route to *nonparametric* identification is to characterize the hazard in some more abstract way. One starting point is to invoke a separability restriction of the sort often made in consumer and producer theory,

$$(5) \quad h(l|x) = m(l|x)\theta.$$

Writing $M(l|x) = \int_0^l h(u|x(u)) du$, we conclude that

$$(6) \quad S(l|x) = \int_{\theta} [\exp - M(l|x)\theta] dG(\theta).$$

In terms of a transformed time scale, $M(L|x)$ is an exponential random variable conditional on θ . The transformed dependent variable has the conditional expectation $E(\ln(M(L|x)|\theta)) = \Gamma'(1) - \ln \theta$, where Γ is the gamma function. Written as a familiar-looking regression equation, the problem of identifying the hand of the past is to determine M and the distribution of Θ from the regression

$$(7) \quad \ln M(L|x) = \Gamma'(1) - \ln \theta + V$$

where $E(V) = 0$, $\text{Var}(V) = \pi^2/6$.³ The severity of the identification problem is evident from (7).

Representation (6) reveals the nature of the available sample information. Level sets of $S(l|x)$, that is, $\{(l, x): S(l|x) = s\}$ trace out level sets of $M(l|x)$. If $M(l|x)$ is differentiable in x , we know the ratio of partial derivatives of M when the right-hand sides

of the following expressions are well defined:

$$(8) \quad \frac{\frac{\partial M(l|x)}{\partial x_i}}{\frac{\partial M(l|x)}{\partial x_j}} = \frac{\frac{\partial S(l|x)}{\partial x_i}}{\frac{\partial S(l|x)}{\partial x_j}},$$

$$\frac{\frac{\partial M(l|x)}{\partial x_i}}{\frac{\partial M(l|x)}{\partial l}} = \frac{\frac{\partial S(l|x)}{\partial x_i}}{\frac{\partial S(l|x)}{\partial l}}$$

where x_i, x_j are elements of x . Assuming $M(0, x) = 0$ for all x (so there are no jumps in the hazard at zero) and $E(\Theta) < \infty$, the right derivative of $S(l|x)$ evaluated in the neighborhood of $l = 0$ is

$$\left. \frac{\partial S(l|x)}{\partial l} \right|_{l=0} = - \left. \frac{\partial M(l|x)}{\partial l} \right|_{l=0} E(\Theta).$$

If $E(\Theta) = 1$ or some other known constant, the magnitude of the local hazard is known for all x .

This information does not suffice to identify M or G . A certain class of monotonic transformations of M explains the data equally well. Denote a member of this class by $J(0) = 0$.

$$(9) \quad S(l|x) = \int_{\theta} [\exp - [M(l|x)]\theta] dG(\theta) \\ = \int_{\theta^*} [\exp - [J(M(l|x))\theta^*] dG^*(\theta^*)$$

where Θ^* is a random variable in general distinct from Θ with distribution G^* .

Observe that $S(l|x)$ is both an exponential mixture of M and an exponential mixture of $J(M)$. As noted by William Feller (1971, p. 452), every mixture of exponentials is an infinitely divisible distribution so by Theorem 1 of Feller (p. 450): $S(M) = e^{-\psi(M)}$ and $S(J) = e^{-\psi^*(J(M))}$ where ψ and ψ^* satisfy the condition that $\psi(0) = \psi^*(0) = 0$ and both have a completely

³See the derivation in Chris Flinn and myself (1982, Appendix B).

monotone first derivative.⁴ Since both $S(M(l|x))$ and $S(J(M(l|x)))$ equal $S(l|x)$ for the same values of l and x , the equivalence class is defined as containing those J such that

$$(10) \quad \psi(M) = \psi^*(J(M)).$$

The requirement that ψ' and $(\psi^*)'$ are completely monotone restricts the admissible J . If $\psi' / (\psi^*)'$, $[(\psi^*)' \neq 0]$ is completely monotonic, then so is J' . Repeated differentiation of (10) produces an algorithmic definition of the derivatives of J . My article with Walker (1990b) uses these results for a class of exponential models ($M(l|x) = l$) and demonstrates the equivalence of a broad class of apparently different duration models within the class of mixture of exponentials models and shows the practical importance of this equivalence for a problem in demography.

Conventional duration models invoke additional separability assumptions:

$$(11) \quad M(l|x) = Z(l)K(x)$$

to produce the "proportional hazards" model. Z is assumed to be nonnegative, differentiable, and monotone increasing in l . $K(x)$ is nonnegative. The previous assumptions ensure that $Z(0) = 0$.

Identification in this class of models has received much attention for the case when X is time invariant. Elbers and Ridder establish that if $E(\Theta) < \infty$ and X assumes at least two distinct values, Z , K , and G are nonparametrically identified. Singer and I (1984) replace the finite mean of Θ assumption with a tail condition on G . Ricardo Barros and Honoré (1988) show that for this model $J(M) = aM^b$, $0 < a$, $0 < b \leq 1$ so J' is completely monotonic. They prove that the components of proportional hazards models are identified up to an arbitrary normalization and up to arbitrary powers (b).⁵

⁴Thus letting $R = \psi'$, R is completely monotone if $(-1)^n R^{(n)}(M) \geq 0$ for all nonnegative integer n and for all $M > 0$ where (n) denotes the n th derivative.

⁵See also the discussion in my book with Singer (1985, p. 64).

(See also Ridder, 1990.) By restricting the mean of Θ or the tail of G , one can avoid this arbitrariness. But such restrictions are themselves arbitrary. Tails of G are not known in advance. Finiteness of the mean for unobservable Θ is not necessarily a reasonable assumption. Models with infinite means for Θ produce entirely reasonable duration models.

Honoré (1990) establishes that if the elements of X are time varying, are not functionally dependent on l , and if one coordinate of X has discrete jumps for a proportion of the population strictly between zero and one and X is time invariant for the rest of the population, then Z , K , and G are nonparametrically identified without invoking arbitrary tail conditions for G or finite mean conditions.

It is interesting to contrast these results with those achieved from an *accelerated hazard* model:

$$(12) \quad M(l|x) = Z(lK(x)).$$

For this model, the assumption $E(\Theta) < \infty$ or the other assumptions do not identify Z or G . However, if $Z'(0) > 0$ and $E(\Theta) < \infty$, and adopting the *normalization* $K(x_*) = 1$, then $K(x)$ is nonparametrically identified up to an arbitrary positive scale.

$$\frac{\frac{\partial S(t|x)}{\partial t}}{\frac{\partial S(t|x_*)}{\partial t}} \bigg|_{t=0} = K(x)$$

We may rewrite (9) using known $q = lK(x)$, to reach $S(q) = \int_0^\infty [\exp - Z(q)\theta] dG(\theta)$. The terms Z , G are not identified except for the special case when $Z(t) = at^b$ (the Weibull) when the proportional hazard model and accelerated hazard model coincide. In that case it is not necessary to have any regressor in order to identify a , b , and G (see myself and Singer, 1984). My article with Honoré notes that $K(x)$ can be identified under weaker assumptions than are required to identify Z and G in both the proportional hazard and accelerated hazard models. It also notes that the accelerated

failure time model is not nonparametrically identified.⁶

Barros (1986) considers the case $M(I|x) = N(Z(I), K(x))$ where Z and K have the properties assumed in the discussion of the proportional hazard model without time-varying variables and where N is assumed known and $N(Z, \cdot) = 0$ if and only if $Z = 0$, N is differentiable with respect to K , $N_K \neq 0$ for $Z \neq 0$ and $N_Z(0, \cdot)$ is finite and different from zero with a well-defined inverse. For $E(\Theta) < \infty$, he proves that Z , K , and G are nonparametrically identified.

III. Conclusion

The ability to distinguish between heterogeneity and duration dependence in single-spell duration models rests critically on maintaining explicit assumptions about the way unobservables and observables interact. A general nonseparable model is nonparametrically underidentified. Separable models are identified subject to additional assumptions on the nature of the explanatory variables (as in Honoré) or subject to restrictions on unobservables. Economically extraneous statistical assumptions drive the answer to the stated question. Viewed as a prototype for identification in general non-ergodic models, these results are not encouraging.⁷

⁶The same qualitative results hold for the more natural version of the accelerated hazard model $Z(tK(x)\theta)$.

⁷Honoré demonstrates how access to multiple spell data facilitates identification of single spell hazards. My article with Honoré discusses conditions for non-parametric identifiability of the competing risks model.

REFERENCES

- Barros, Ricardo, "The Identifiability of Hazard Functions," working paper, University of Chicago, 1986.
- _____ and Honoré, Bo, "Identification of Duration Models with Unobserved Heterogeneity," working paper, Northwestern University, 1988.
- Elbers, Chris and Ridder, Geert, "True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model," *Review of Economic Studies*, July 1982, 49, 403-09.
- Feller, William, *An Introduction to Probability Theory and Its Applications*, Vol. II, New York: Wiley & Sons, 1971.
- Flinn, Chris and Heckman, James, "Models for the Analysis of Labor Force Dynamics," in G. Rhodes and R. Bassman, eds., *Advances in Econometrics*, Vol. 1, 1982, 35-95.
- Heckman, James and Honoré, Bo, "The Identifiability of the Competing Risks Model," *Biometrika*, June 1989, 76, 325-30.
- _____ and Singer, Burton, "The Identifiability of the Proportional Hazard Model," *Review of Economics Studies*, July 1984, 51, 231-43.
- _____ and _____, "Social Science Duration Analysis" in their *Longitudinal Analysis of Labor Market Data*, Cambridge: Cambridge University Press, 1985, ch. 2.
- _____ and Walker, James, (1990a) "The Relationship Between Wages and Income and The Timing and Spacing of Births: Evidence From Swedish Longitudinal Data," *Econometrica*, November 1990, 58, 1411-42.
- _____ and _____, (1990b) "Estimating Fecundability from Data on Waiting Times to First Conception," *Journal of the American Statistical Association*, June 1990, 85, 283-94.
- Honoré, Bo, "Identification Results for Duration Models with Multiple Spells or Time-Varying Covariates," unpublished manuscript, Northwestern University, 1990.
- Lancaster, Tony, "Econometric Methods for the Duration of Unemployment," *Econometrica*, September 1979, 47, 939-56.
- Ridder, Geert, "The Nonparametric Identification of Generalized Accelerated Failure-Time Models," *Review of Economic Studies*, May 1990, 57, 167-81.
- Silcock, Herbert, "The Phenomenon of Labor Turnover," *Journal of the Royal Statistical Society, Series A*, July 1954, 117, 429-40.

History and Industry Location: The Case of the Manufacturing Belt

By PAUL KRUGMAN*

If there is one single area of economics in which path dependence is unmistakable, it is in *economic geography*—the location of production in space. The long shadow cast by history over location is apparent at all scales, from the smallest to the largest—from the cluster of costume jewelry firms in Providence to the concentration of 60 million people in the Northeast Corridor.

This paper illustrates path dependence in economic geography by describing a particular historical example, the persistence of the U.S. “manufacturing belt,” and a simple model that helps make sense of that example.

I. The U.S. Manufacturing Belt

Early in this century, geographers noted that the great bulk of U.S. manufacturing was concentrated in a relatively small part of the Northeast and the eastern part of the Midwest. This manufacturing belt took shape in the second half of the nineteenth century, and proved remarkably persistent. Harvey Perloff (1960) estimated that as late as 1957 the manufacturing belt still contained 64 percent of U.S. manufacturing employment—only slightly reduced from its 74 percent share at the turn of the century.

Even this number understates the dominance of the belt, because during its heyday most manufacturing outside it consisted either of processing of primary products or of production for a very local market. That is, the manufacturing belt contained virtually all manufacturing that did not need either to be close to the consumer or close to specific natural resources.

The manufacturing belt persisted even as the center of gravity of agricultural and

mineral production shifted far to the West. In 1870, the Northeast and East North Central regions (within which the emerging manufacturing belt lay) accounted for 44 percent of U.S. “resource extraction” employment (agriculture, mining, forestry, fisheries). By 1910, this share had already fallen to 27 percent; yet these regions still accounted for 70 percent of manufacturing employment. And whereas the belt’s share of manufacturing employment understates its manufacturing dominance, its share of resource employment overstates its resource base, since much of the agriculture in or near the manufacturing belt took place less because of the suitability of the land than because of proximity to urban centers.

The manufacturing belt’s persistent dominance evidently reflects some kind of external economies. But what was the nature of these external economies? I will sketch out a simple model that surely does not capture the full story, but is strongly suggestive of the kind of explanation that is needed.

II. A Core-Periphery Model¹

A core-periphery pattern like that of industrial America can emerge from the interaction of increasing returns, transportation costs, and demand. Given sufficiently strong economies of scale, each manufacturer wants to serve the national market from a single location. To minimize transportation costs, she chooses a location with large local

¹This paper presents only a sketch of a model. It will be apparent that this sketch is sloppy about a number of issues, including: what is the market structure in manufacturing; what happens to profits, if any; and what resources are used in both fixed costs and transportation. It is possible to derive similar results in a fully specified general equilibrium monopolistic competition model (see my 1991 article). I adopt the more *ad hoc* approach here for ease of exposition.

*Department of Economics, MIT, Cambridge, MA 02139.

demand. But local demand will be largely precisely where the majority of manufacturers choose to locate. Thus there is a circularity that tends to keep a manufacturing core in existence once it is established. (This is not an original story: it is more or less explicit in the work of such geographers as Allan Pred, 1966, and David Myers, 1983.)

Imagine a country in which there are only two possible locations of production, East and West, and two kinds of production. Agricultural goods are produced using a location-specific factor (land), and as a result the agricultural population is exogenously divided between the locations; we assume that the division is 50–50.

Manufactured goods (of which there are many symmetric varieties) can be produced in either or both locations. If a given manufactured good is produced in only one location, transportation costs must be incurred to service the other market. On the other hand, if the good is to be produced in both locations, an additional fixed set-up cost is incurred. The manufacturing labor force in each location is proportional to manufacturing production in that location.

Finally, assume that the *demand* for each manufactured good in each location is strictly proportional to that location's population.

The basic workings of the model can then be illustrated by Figure 1. On the horizontal axis we measure the share of the manufacturing labor force employed in West, on the vertical axis the share of West in the total population. The line *MM* represents the dependence of the distribution of manufacturing on the distribution of population; the line *PP* the converse effect of manufacturing on population distribution.

If West has a small share of the population, it will not be worth incurring the fixed costs of establishing a manufacturing facility there; it is cheaper to serve the market from facilities in East. Conversely, if West has a large share of the population, it is not worth producing manufacturers in East. A sufficiently equal division of population might, however, lead manufacturers to produce locally for both markets. Putting these observations together, we get the illustrated shape

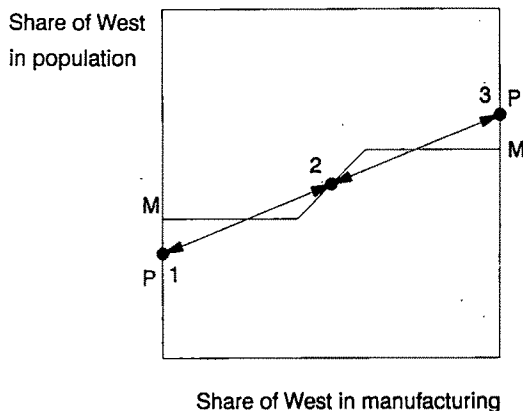


FIGURE 1

of *MM*: no western production for low western population, production proportional to population for intermediate levels, no eastern production if the West is big enough.

The more manufacturing that takes place in West, the larger the western share of the total population; but not in proportion, because some people are immobile farmers. Thus *PP* is upward sloping, but flatter than the 45° line.

Suppose that manufacturing production adjusts gradually toward its equilibrium level. Then the dynamics are illustrated by the arrows in Figure 1. There are three stable equilibria: manufacturing may be concentrated in either location, at 1 or 3, or it may be equally divided, at 2.

The picture need not look like this: a core-periphery geography may be unsustainable because the “centripetal” pull of a manufacturing core is weaker than the “centrifugal” pull of the dispersed agricultural sector. We can easily derive a necessary condition for concentration of manufacturing production in one location. Consider a typical manufacturing firm with total sales of x ; let π be the proportion of the population that is employed in manufactures; let F be the fixed cost that the firm must incur to produce in both locations; and let t be the transportation cost per unit shipped between locations. Now we ask: is a situation with all manufacturing concentrated in East an equilibrium?

With all manufacturing in East, West has a share of total population equal to only $(1 - \pi)/2$. The transportation cost of serving this market from East for a typical manufacturer is therefore $tx(1 - \pi)/2$. The cost of setting up a plant in West is F . So a concentration of production in East, once established, will persist as long as $F > tx(1 - \pi)/2$. If this criterion is not met, there is no path dependence: the long-run geography of manufactures will follow that of agriculture.²

We can immediately see that whether geography is path dependent is determined by three parameters: large F , that is, sufficiently strong economies of scale; small t , that is, sufficiently low costs of transportation; and large π , that is, a sufficiently large share of "footloose" production not tied down by natural resources.

We can now tell a stylized story of the emergence of the manufacturing belt. In the early United States, with its primarily agricultural population, where manufacturing was marked by few scale economies, and where transportation was costly, no strong geographical concentration could occur. As the country began its industrial transition, manufacturing arose in areas that contained most of the agricultural population outside the South. During the second half of the nineteenth century, however, manufacturing economies of scale increased, transportation costs fell, and the share of the population in nonagricultural occupations rose. The result was that the initial advantage of the manufacturing belt was locked in. Even though new land and new resources were exploited to the West and even though slavery ended, for three-quarters of a century the pull of the established manufactured areas was strong enough to keep the manufacturing core virtually intact.

Of course this story oversimplifies. In particular, it says nothing about the sources of local specialization within the manufactur-

ing belt—about why Detroit emerged as the automotive center, New York as the garment center, Grand Rapids as the furniture center, etc. Yet it surely captures an important aspect of what happened. And it also contains elements (increasing returns at the level of individual firms, and external economies resulting from the interaction of these firms' decisions) that will reappear as one further elaborates the story.

III. Further Thoughts

The case of the U.S. manufacturing belt is of substantial interest in its own right. The rise and persistence of that belt is an important yet much neglected aspect of U.S. economic history. More important than its immediate significance, however, is what the history of manufacturing location says about the nature of our economy in general. And what it says is Nicholas Kaldor (1972), Paul David (1985), and Brian Arthur (1989) were right—that increasing returns and cumulative processes are pervasive and give an often decisive role to historical accident.

It is also interesting that the story of the manufacturing belt reaches back to the mid-nineteenth century. It is common to argue that external economies and cumulative processes have become more important in recent decades because of the growing importance of high technology. The geographical concentration of manufacturing in the United States took shape, however, in the age of steam, not that of microprocessors. So it is not simply true that our economy is not now well described by the conventional constant returns model. It never was.

REFERENCES

- Arthur, Brian, "Positive Feedbacks in the Economy," *Scientific American*, February 1989, 262, 92–99.
- David, Paul, "Clio and the Economics of QWERTY," *American Economic Review Proceedings*, May 1985, 75, 332–37.
- Kaldor, Nicholas, "The Irrelevance of Equilibrium Economics," *Economic Journal*,

²This is a necessary condition for concentration. A sufficient condition for concentration of manufacturers' production in one location is $F > tx/2$. In this case, the 45° section of *MM* disappears, and it becomes a horizontal line at an equal population division.

December 1972, 82, 1237-55.

Krugman, Paul, "Increasing Returns and Economic Geography," *Journal of Political Economy*, forthcoming 1991.

Myers, David, "Emergence of the American Manufacturing Belt: An Interpretation," *Journal of Historical Geography*, April

1983, 9, 145-74.

Perloff, Harvey, *Regions, Resources, and Economic Growth*, Baltimore: Johns Hopkins University Press, 1960.

Pred, Allan, *The Spatial Dynamics of U.S. Urban-Industrial Growth, 1800-1914*, Cambridge: MIT Press, 1966.

Complementarities, Momentum, and the Evolution of Modern Manufacturing

By PAUL MILGROM, YINGYI QIAN, AND JOHN ROBERTS*

In the nineteenth century, the railroad and telegraph were at the center of a set of technological advances, physical investments, and managerial innovations that transformed American industry (Alfred Chandler, 1977). Later, the automobile and telephone played a similar role in another transformation.

Today, the high-tech industries include computers, telecommunications, and electronics. Working on our remarkably powerful computers (even as they rapidly become obsolete), co-authoring papers by electronic mail and fax, and conversing on our portable cellular telephones, we are struck by what appears to be a self-supporting and -reinforcing dynamic to the technological improvements across the electronics industries. An advance almost anywhere in the sector seems to call forth more advances across the sector.

These advances are occurring contemporaneously with a broad pattern of other changes, not only in the electronics industries, but in manufacturing more generally, and not just in hardware, but in methods and organization as well. A new paradigm has begun to emerge. In contrast to traditional manufacturing firms, modern firms frequently 1) make greater use of flexible, programmable equipment and of computer-aided design and manufacturing technologies, 2) have fewer job classifications, 3) offer more varieties of their major products and/or update their product lines more frequently, 4) put more emphasis on speed in

order processing, production, and delivery, 5) hold much lower inventories of intermediate and finished goods, 6) rely on subcontractors to supply a greater proportion of the total value added, and 7) overlap design, product, and process engineering to speed the introduction of new products. These features of modern manufacturing firms encompass technology choices, marketing strategies, personnel policies, supplier relations, lines of internal communications, and other operational policies in a far-reaching and coherent pattern whose existence, in the words of Michael Piore (1986), poses a "challenge to economic theory."

In a recent paper, Milgrom and Roberts (1990), proposed a theory to explain the emergence of this new paradigm, arguing that the various characteristics and activities described are mutually complementary and so tend to be adopted together, with each making the others more attractive. In that theory, the falling costs of high-speed data communication, data processing, and flexible, multitask equipment lead to increases in the directly affected activities, which through a web of complementarities then lead to increases in a set of related activities as well.

Although the costs of flexible machinery and data communication and processing surely have fallen substantially in recent decades, an analysis that takes these changes as exogenous does not constitute a full explanation of the pattern of change that we observe. One must ask: Why are these particular costs falling relative to others? Here we enrich the Milgrom-Roberts analysis by adding a dynamic to it: innovations in the manufacture of basic inputs both arise in response to a growing market for those inputs and simultaneously encourage that growth. This reflects a fundamental complementarity between the level of any activity and investments that reduce its marginal

*Department of Economics (Milgrom and Qian) and Graduate School of Business (Roberts), Stanford University, Stanford, CA 94305. We thank Avner Greif for his helpful comments on an earlier draft. Financial support from the National Science Foundation (Milgrom and Roberts) and the Robert and Anne Bass Faculty Fellowship (Roberts) is gratefully acknowledged.

cost. In our model, innovation results in falling marginal costs that lead to increasing usage of the inputs, which in turn leads to increasing investments in the development of complementary techniques, that further raise the underlying demand for the inputs, leading to more innovations, and so on, just as we have noted within the electronics industry itself.

Significantly, just as this verbal sketch requires no assumptions about the presence or absence of returns to scale, neither does our formal analysis: The momentum of the system of changes we analyze results entirely from the positive feedback effects that each of a group of core activities and practices has on the *other* activities and practices in the group.¹

Our model also adds to the Milgrom-Roberts model groups of "additional" activities, each of which may interact with any one core activity in a general way but that are not themselves part of the *mutually complementary core group*.

Our formal analysis is summarized in two theorems. Most important is the *Momentum Theorem*: this asserts that once the system begins along a path of growth of the core variables, it will continue forever along that path or, more realistically, until unmodeled forces disturb the system. Second, the *Reduced-Forms Theorem* concerns the set of models that have a reduced form to which the Momentum Theorem applies.

I. The Downstream Industry Model and the Reduced-Forms Theorem

We suppose that there is a representative firm in the "downstream manufacturing industry" whose profit function in period t depends on the current state of (public) knowledge, represented by a vector $\theta(t) \in \mathbb{R}^k$, on a vector of core decision variables

$x = (x^1, x^0) \in \mathbb{R}^{m+n}$, where the variables $x^1 = (x_1, \dots, x_m)$ denote purchased *inputs* and the variables $x^0 = (x_{m+1}, \dots, x_{m+n})$ denote the *other* core decision variables, and finally on various variables outside the core group and denoted by $\{y^i(t); 1 \leq i \leq m+n\}$. The firm's payoff takes the form:

$$\rho(x(t), \theta(t)) + \sum_j \phi_j(x_j(t), y^j(t)) - R(x^1(t)),$$

where $R(x^1(t))$ is the amount paid for inputs. If ρ is smooth, the assumptions that the components of x are mutually complementary and that their effectiveness is enhanced by technical knowledge are represented by the inequalities $\partial^2 \rho / \partial x_i \partial x_j \geq 0$ for all $i \neq j$ and $\partial^2 \rho / \partial x_i \partial \theta_j \geq 0$ for all i and j , where $1 \leq i, j \leq m+n$. That is, the marginal product of any component of x is nondecreasing in the levels of the other arguments and in the level of technological knowledge. More generally, the first assumption is that ρ is *supermodular*, that is, for all x, x' , and θ ,

$$\rho(x, \theta) + \rho(x', \theta) \leq \rho(\text{Max}(x, x'), \theta) + \rho(\text{Min}(x, x'), \theta),$$

where Max and Min are taken component-by-component. The second is that ρ has *increasing differences*, that is, if for all $x \geq x'$, the difference $\rho(x, \theta) - \rho(x', \theta)$ is nondecreasing in θ . We emphasize that there are no assumptions made about the own second partial derivatives, $\partial^2 \rho / \partial x_i^2$, so that the profit function may be convex in x_i over some ranges and concave over others. No assumptions are made about returns to scale.

The y^i 's represent other variables, each of which interacts with at most one core variable. For example, if one of the core decision variables x_i is the level of inventories to hold, the y^i vector might include decisions about where to hold the inventory or how large the storage room should be. The restriction is that none of these should affect the returns to other core variables, such

¹ Indeed, our entire analysis is a purely ordinal one, that is, no step in the argument is affected if the variables are subject to arbitrary monotone rescalings. Since returns to scale is a cardinal concept, all of our conclusions apply without any assumptions about returns to scale.

as the flexibility of equipment or the number of job classifications.

We assume that each decision variable x_i or y^i is constrained to lie in some compact set $S(x_i)$ or $S(y^i)$, which may be finite or infinite. Our main conclusion about these noncore variables is that their presence has no qualitative effect on the analysis.

THEOREM 1. Reduced Forms: Suppose ρ and each ϕ_j ($j = 1, \dots, m + n$) is continuous. Then the reduced-form (gross) profit function of the manufacturing sector given by

$$(1) \quad \pi(x, \theta) = \text{Max}_y \rho(x, \theta) + \sum_j \phi_j(x_j, y^j),$$

is continuous, supermodular in x , and has increasing differences in (x, θ) .

PROOF:

Let $G(x) \equiv \text{Max}_y \sum_j \phi_j(x_j, y^j)$. Then, because of the separability of the y variables, $G(z) + G(z') = G(\text{Max}(z, z')) + G(\text{Min}(z, z'))$. So, for all θ , $[\pi(z, \theta) + \pi(z', \theta)] - [\pi(\text{Max}(z, z'), \theta) + \pi(\text{Min}(z, z'), \theta)] = [\rho(z, \theta) + \rho(z', \theta)] + [G(z) + G(z')] - [\rho(\text{Max}(z, z'), \theta) + \rho(\text{Min}(z, z'), \theta)] - [G(\text{Max}(z, z')) + G(\text{Min}(z, z'))] = [\rho(z, \theta) + \rho(z', \theta)] - [\rho(\text{Max}(z, z'), \theta) + \rho(\text{Min}(z, z'), \theta)] \leq 0$. The last inequality holds because $\rho(x, \theta)$ is supermodular in x for any given θ . Finally, from $\pi(z, \theta) - \pi(z', \theta) = \rho(z, \theta) - \rho(z', \theta) + G(z) - G(z')$, we conclude that $\pi(x, \theta)$ should also have increasing differences in (x, θ) since $\rho(x, \theta)$ has increasing differences in (x, θ) .

II. The Upstream Industry and the Contracting Equilibrium

We also suppose that there is a representative firm in the "upstream," input-producing sector. Its profit function is $R(x^I(t)) - C(x^I(t), T(t), \eta(t))$, where $\eta(t)$ is the "knowledge" vector describing knowhow in the input industry in period t and $T(t)$ is the level of technology it uses in producing its output. We assume that $-C$ is supermodular in (x, T) and has increasing differences in (x, T) and η . Thus, the core aspects of the technology are separable or

complementary ($\partial^2 C / \partial T_k \partial T_j \leq 0$) and there are no diseconomies of scope in producing the various inputs ($\partial^2 C / \partial x_i^I \partial x_j^I \leq 0$). Also, increases in both public knowledge $\eta(t)$ and the firm's own technology ($T(t)$) reduce its marginal cost of production: $\partial^2 C / \partial x_j^I \partial T_k \leq 0$ and $\partial^2 C / \partial x_j^I \partial \eta_k \leq 0$ for all j and k . Finally, accumulated public knowledge is assumed to reduce the marginal cost of technological improvements: $\partial^2 C / \partial T_j \partial \eta_k \leq 0$.²

Given the limited space, we set aside the model of consumers and questions of the nature and existence of equilibrium. Rather, we assume that the upstream firms and their suppliers arrange terms that are efficient for themselves, ignoring whatever effects their activities may have on the economywide accumulation of knowledge. This latter assumption (that firms ignore the effect of their efforts on public knowledge) is indicative of a free-rider problem and is most reasonable when the typical firm is not too large. Then, the industry equilibrium involves choosing x and T to maximize the objective:

$$(2) \quad \pi(x(t), \theta(t)) - C(x^I(t), T(t), \eta(t)),$$

where π is the reduced-form profit function given in (1).

III. Knowledge Accumulation Dynamics and the Momentum Theorem

The knowledge $\theta(t)$ and $\eta(t)$ at date t is assumed to be freely available to all firms. Suppose that $\theta(t+1) = f(\theta(t), \eta(t), x(t), T(t))$, and $\eta(t+1) = g(\theta(t), \eta(t), x(t), T(t))$, where both f and g are nondecreasing functions. That is, higher levels of the core activities in the downstream sector and higher levels of technology in the upstream sector combine with a higher initial

²Generally, using high-technology equipment may require both "knowledge" $\eta(t)$ and cash investment $I(t)$, for example: $T(t) = \eta(t)h(I(t))$, where $h' > 0$ and $h'' < 0$. In this example, the cost of achieving technology level T using knowledge η is $I = h^{-1}(T/\eta)$, which implies the required inequality: $\partial^2 I / \partial T \partial \eta < 0$.

state of knowledge today to produce a higher state of knowledge tomorrow. The model incorporates the possibility that learning by doing and high levels of activity in one industry increase learning in the other.

THEOREM 2. Momentum: *Suppose that for every given value of (θ, η) , there is a unique (x, T) that maximizes (2). If there is any time t such that $\theta(t) \geq \theta(t-1)$ and $\eta(t) \geq \eta(t-1)$, then for all times $s \geq t$, $x(s) \geq x(s-1)$, $T(s) \geq T(s-1)$, $\theta(s) \geq \theta(s-1)$ and $\eta(s) \geq \eta(s-1)$.*

PROOF:

By Theorem 1, $\pi(x, \theta) - C(x^I, T, \eta)$ is supermodular in (x, T) and has increasing differences in $(x, T; \theta, \eta)$. Therefore, given our hypothesis that $\theta(t) \geq \theta(t-1)$ and $\eta(t) \geq \eta(t-1)$, we conclude that $x(t) \geq x(t-1)$ and $T(t) \geq T(t-1)$ according to the Topkis's monotonicity theorem (Donald Topkis, 1978).³ $\theta(t+1) \geq \theta(t)$ and $\eta(t+1) \geq \eta(t)$ then follow immediately from the fact that f and g are nondecreasing. The theorem then follows by induction.

Although we have not explored it here, our model does allow the possibility of multiple steady states. Nevertheless, because our formal model has no durable capital in the firms and because our representative firms always contract efficiently, the possibility that expectations affect the path of the economy is excluded. Instead, following Paul David (1988) and Steven Durlauf (1990), we emphasize the role of history, featuring the momentum of the economic system.

IV. Conclusion

The Momentum Theorem shows that complementarities among a group of core activities and processes can account for the emergence of a persistent pattern of change,

even without any of the usual assumptions in the growth literature about economies of scale (see Paul Romer, 1986). Our method, emphasizing complementarities over issues of scale, also promises to clarify the logic of growth models and to allow a far richer modeling of the multifaceted processes of growth and development.

In Chandler's account of nineteenth-century American economic growth, for example, the emergence of the large industrial enterprise was accompanied not only by improvements in communications (telegraph) and transportation (railroads) that helped to create national markets, but also by innovations in finance (bond markets), management methods (cost accounting), large-scale manufacturing technologies (continuous process technologies), and so on. Each of these improvements and innovations were complementary to further growth of large enterprises, and the expanding scale of these enterprises correspondingly encouraged continuing technological, organizational and managerial advances.

Closely related to our ideas are some longstanding analyses of economic development, where the need to manage complementarities among investment projects has been noted by some economists (see Albert Hirschman, 1960). The questions addressed in these analyses are not merely ones of whether to develop or how to develop, but also *in which direction* to develop. An analysis of complementarities, richly conceived, seems indispensable to giving a satisfactory answer to these questions.

REFERENCES

- Chandler, Alfred D., *The Visible Hand: The Managerial Revolution in American Business*, Cambridge: Harvard University Press, 1977.
- David, Paul A., "Path-Dependence: Putting the Past into the Future of Economics," IMSSS Technical Report No. 533, Stanford University, 1988.
- Durlauf, Steven, "Nonergodic Economic Growth," working paper, Stanford University, 1990.

³Topkis's original treatment is very general and abstract. A simpler account of Topkis's Theorem, restricted to applications in \mathbb{R}^N , can be found in the paper by Milgrom and Roberts.

- Hirschman, Albert, O., *The Strategy of Economic Development*, New Haven: Yale University Press, 1960.
- Milgrom, Paul and Roberts, John, "The Economics of Modern Manufacturing: Technology, Strategy, and Organization," *American Economic Review*, June 1990, 80, 511-28.
- Piore, Michael J., "Corporate Reform in American Manufacturing and the Challenge to Economic Theory," mimeo., MIT, 1986.
- Romer, Paul M., "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, October 1986, 94, 1002-37.
- Topkis, Donald M., "Minimizing a Submodular Function on a Lattice," *Operations Research*, March-April 1978, 26, 305-21.

PRICE STICKINESS IN THEORY AND PRACTICE

Why are Prices Sticky? Preliminary Results from an Interview Study

By ALAN S. BLINDER *

Those of us who teach Keynesian economics (an increasing number these days) attach great importance to the phenomenon of wage and/or price "stickiness." It explains, for example, why recessions cure themselves only slowly and why changes in the money supply have real effects. In fact, Keynesian economics is sometimes characterized as the economics of nominal rigidities. Yet, more than a half-century after Keynes published *The General Theory*, the phenomenon itself remains poorly understood. Just why are wages or prices sticky?

It is not that economists have ignored these questions. One could literally fill many volumes with good empirical studies of wage and price stickiness, and many more with clever theories purporting to explain these phenomena. Yet, despite all this work, the range of admissible theories is wider than ever, and new theories continue to crop up faster than old ones are rejected. This lack of scientific progress makes one wonder

about the basic research strategy that economists have been pursuing. Is there a better way?

In Section I, I argue that traditional research strategies may be unpromising vehicles for learning about why prices are sticky; the time may have come to entertain new and unorthodox approaches. Section II defends the notation that interviews, in particular, can be a useful research tool in this area. Sections III and IV then describe the design of and summarize some preliminary findings from a large-scale interview study currently underway at Princeton. This paper is very much a progress report since the interviewing still has 9–12 months to go.

I. Why Interviews?

The standard program of scientific research in economics is to (a) develop a theory of some phenomenon, (b) formulate it in econometric terms, and then (c) test it with actual data. The theory is then either rejected or provisionally accepted, that is, allowed to survive to the next test. Unfortunately, this program has been singularly unsuccessful in the area of wage-price stickiness. Most economists would, I think, agree that we know next to nothing about which of several dozen theories of wage-price stickiness are valid and which are not. We might have expected statistical tests to have weeded out the weaker theories by now, especially since many of them have been around a long time. But, in fact, the survivorship rate rivals that of congressional elections. Why?

Part of the reason, no doubt, is that economists are fonder of building theories than of testing them. But I think the main reason is that most of the theories are *empty*

*Department of Economics, Princeton University, Princeton, NJ 08544. I owe many debts of gratitude: to the Russell Sage and Alfred P. Sloan Foundations for financial support; to Eric Wanner for persuading me to go ahead with the project; to Elie Canetti for extensive and excellent research assistance; to Phyllis Durepos for efficiently handling the considerable volume of office traffic; to my graduate student interviewers, Anthony Marcus, David Genesove, Katy Graddy, Dean Jolliffe, Harold Kim, John Leahy, Alec Levenson, John Penrod, Michael Quinn, Steve Schwartz, Tim Vogelsang, and David Zimmerman; to Bob Abelson, George Akerlof, Larry Ball, Olivier Blanchard, Elizabeth Bogan, Dennis Carlton, Russell Cooper, Bob Gordon, Bruce Greenwald, Glenn Hubbard, Tom Juster, Danny Kahneman, Anil Kayshap, Alan Krueger, David Romer, Julio Rotemberg, Bob Shiller, and Andy Weiss for helpful suggestions; and, especially, to the many business executives who took time out from their busy schedules to answer our questions.

in the following specific sense: Either they involve unobservable variables in an essential way, or they carry no real implications other than that prices are "sluggish" in some unmeasurable sense, or both. This makes econometric modeling a blunt, perhaps even useless, investigative tool. Let me illustrate.

To begin with, think about what we mean when we say that a theory predicts that prices are "sticky": Often nothing more than that prices adjust less rapidly than Walrasian market-clearing prices. But since equilibrium price movements often go unmeasured, it is hard to know whether actual prices are moving faster or slower than this norm. More important, all the theories share *exactly the same* prediction: that prices are sticky in this sense. So how are we to discriminate among them? It seems difficult even to imagine what a decisive test would look like, much less to carry one out.

A natural idea is to use each theory to derive other, auxiliary predictions, and then test these. Unfortunately, often there are no such predictions—or at least none that can be checked against actual data. One reason is that many of the theories are based on variables that are *unobservable* either in principle or in practice.

As an example of the former, consider the theory that firms hesitate to cut prices because they fear that customers will interpret a price cut as signal that quality has been reduced—*when, in fact, there has been no quality reduction*. Clearly, the theory is predicated on the existence of *unobservable* quality differences for, if consumers could observe quality readily, price would not play a signalling role. As an example of the latter, consider the so-called menu cost theory. In principle, fixed costs of changing prices can be observed and measured. In practice, such costs take disparate forms in different firms, and we have no data on their magnitude. So the theory can be tested at best indirectly, at worst not at all.¹

¹For example, the menu cost theory predicts firms will never make "small" price changes, which appears to be a testable proposition. But does it rule out price increases of 1 percent or 10 percent? Without measuring the menu costs, we cannot answer this question.

One might argue that theories that make identical predictions are, in an operational sense, the same theory; so it does not matter which theory is correct. In fact, however, it does matter. It matters for the value of economics as a descriptive science; and it matters for the conduct of macroeconomic policy because not all sources of price rigidity open the door to welfare-improving policies.

If it really matters which theory is correct, but conventional modes of economic inquiry cannot adjudicate the dispute, economic science would appear to be in deep trouble. Fortunately, one other common characteristic of the theories suggests a way out: Virtually every theory of price rigidity *describes a chain of reasoning which allegedly leads the firm to conclude that a change in price is inadvisable*.

That is what gave me the idea for an interview study. If actual decision makers really think the way one of these theories says, they ought to know that they do. If you ask them open-ended questions like, "Why don't you cut your prices when sales decline?," you may get shrugs or incoherent answers. But, if you confront them with the chain of reasoning they actually follow, they ought to recognize and agree with it. Conversely, if they explicitly deny the relevance or validity of a particular argument, then it is probably not governing their behavior. At least that is my methodological precept.

II. But Aren't Interviews Unreliable?

Economists are skeptical that you can learn much by asking people. We are trained to study behavior by watching what people *do* (usually in markets), not by listening to what they *say*. For example, critics will point out that subjects of interviews have no incentive to respond truthfully or thoughtfully; so *homo economicus* might refuse to cooperate or even give misleading answers. If the respondent has reason to conceal the truth or mislead the interviewer, this objection is probably a show stopper. But, in the case of price stickiness, people have no particular reason to conceal the truth. As long as people are not pathological liars, interviews may elicit useful information.

The thoughtfulness problem goes deeper. We all know the billiard ball analogy: A good pool player makes excellent use of the laws of physics without understanding them, and certainly without being able to articulate them. For this reason, many economists doubt that much can be learned by asking "economic players" about how they play the game. But I believe that more pointed questions, posed in plain English, can elicit useful answers. For example, if you ask a skilled billiards player whether he bases his shots on the principle that the angle of incidence equals the angle of reflection, he will probably look at you quizzically. But, if you take him to the table and, pointing to the proper angles, ask: "Do you try to make this angle the same as that angle?," I imagine he would respond in the affirmative.

Legitimate questions can also be raised about the size and representativeness of interview samples. Detailed case studies of two or three companies can, at most, provide anecdotes, not useful statistical information. And samples that are unrepresentative of the underlying population give no basis for drawing inferences about population statistics. But these are familiar problems, well known to any user of data. They are not reasons to reject interview evidence out of hand.

We should remember that theory and econometrics have their limitations, too. Theoretical deductions are often untested or, worse yet, *untestable*. Econometric evidence is often equivocal and/or subject to methodological dispute. The imperfect knowledge we can pick up from interviews and questionnaires should therefore not be compared to some epistemological ideal, but to the imperfect knowledge that nonexperimental scientists can deduce theoretically or glean from econometric studies. By this more reasonable standard of evidence, data culled from interviews certainly look admissible.

III. The Research Design

Along with a team of Princeton graduate students, I have now been in the business of interviewing executives about their pricing strategies since August 1988. Let me briefly

describe the research strategy and some of the tactics.

Designing the Questionnaire: The first step, obviously, was to decide which theories were to be tested and then to turn them into questions that ordinary people can understand. This selection process was not entirely objective; nor could it have been. In scouring the intellectual waterfront, I excluded theories that sounded silly or that actually *were* silly, or that seemed too complicated to be explained tersely in plain English. These omissions were not particularly costly; they did not disqualify any of the major competing theories.

Translating the equations and diagrams into English proved to be easy. My first stab at a questionnaire was tried out on many "guinea pigs" (including economists, business people, and a few scholars involved in survey research) and altered in literally hundreds of ways. But, in general, the current version of the questionnaire bears a striking resemblance to my original draft. Translating from technical journalese into English simply proved not to be very difficult. I took this to be a good omen.

However, one important caveat should be entered. Not infrequently, a question must be rephrased on the spot to make the respondent understand it. To do so successfully, an interviewer must be reasonably articulate, must be able to think on his or her feet, and, most importantly, must understand the economics well enough to paraphrase a question without changing its meaning. Thus professional pollsters will not do. Instead, my interviewers are carefully selected Princeton graduate students.

Eliciting Cooperation from Businesses: The next problem was getting our feet in the door and getting our questions answered by the right people. A low-response rate would obviously raise fears of selectivity bias.

A small-scale pilot study was used both to estimate the likely response rate in a large-scale study and to polish our questionnaire and interviewing techniques. We randomly selected 16 firms in the northeastern United States and wrote each an introductory letter requesting an interview. We then followed up with phone calls and/or further mail-

ings as necessary and, after considerable effort, successfully interviewed 8 companies. (These 8 companies are *not* included in the data reported below.) The estimated 50 percent response rate struck me as high enough to merit proceeding to the full study. So far, the response rate in the full study is running above 60 percent.

In addition, we had no trouble making connections with the person or persons in each company who could answer our questions. The number of "I don't know" responses has been small, and almost never do we hear, "You're asking the wrong person." Once in the door, we have found people more than willing to talk frankly. They find the questions interesting, mostly understandable, and not invasive of privacy.² Our experience, in fact, is that people are often eager to tell us things about which we would never have dared ask for fear that we would appear to be prying. We do, of course, promise confidentiality.

Getting Economists to Pay Attention: The hardest problem may yet lie ahead: Getting economists to pay attention to the results. Several concrete steps have been taken to increase the believability of the ultimate findings, whatever they may be. First, I wrote letters to the originators or major proponents of each theory. The letters explained the nature of the study, included a copy of the relevant portion of the questionnaire, asked for suggested improvements in the questions, and asked the "theorists" to suggest other testable implications of their theories. Most responded, sometimes offering useful ideas for modifying the questionnaire, which were adopted. Interestingly, however, not one person suggested a single further implication that could be tested in the questionnaire. This underscores the point I made earlier about the theories being empty.

Second, and much more important, we have taken great pains to ensure that the sample of firms is (a) large enough for seri-

ous statistical analysis, (b) randomly selected, and (c) representative of the private, for-profit GNP. To my knowledge, this is the first time anything like this has ever been attempted in an interview study. Since it is what makes the study unique, it is worth describing in some detail.

We purchased a tape listing firms with annual sales over \$10 million in the northeastern United States, an area which accounts for 40–45 percent of U.S. GNP. From this sampling universe, we eliminated government enterprises and nonprofits on the grounds that the theories we wanted to test were all about profit-seeking firms; this left almost 25,000 firms. We then assigned a sampling weight to each firm proportional to its value-added and drew a random sample of 400 firms, seeking to complete 200 interviews. (Remember, the estimated response rate was 50 percent.)

Notice that we excluded any company with annual sales under \$10 million, even though they employ nearly half the labor force. Why? Because there are so many of them, and the expense of reaching any sizable portion would be prohibitive. Clearly, an optimal experimental design would balance the value of the information obtained against the costs of obtaining it and would, therefore, undersample very small companies.³ We approximated this crudely by not sampling any companies below the \$10 million threshold. Analogous cost considerations motivated the geographical truncation: costs depend on distance from Princeton, NJ; benefits do not.

These two exclusions, however, might compromise the representativeness of the sample. So we remedied them as follows. Suppose pricing behavior varies geographically only because industrial structure differs across states, not because, say, California firms are inherently different from New Jersey firms. Then we can (and did) create a *synthetic national sample* by reweighting each firm in the Northeast to reflect na-

²Considerable effort was expended to guarantee the latter. For example, I abjured any questions that might conjure up images of the U.S. Department of Justice.

³This presumes that the study aims to explain the behavior of the GNP deflator.

tional, rather than regional, shares in value-added. Firms in industries that are overrepresented in the Northeast, like banking, were appropriately undersampled; and firms in industries underrepresented in the region, like gas and oil extraction, were correspondingly oversampled. Thus, the industrial structure of our sampling frame matches that of the nation as a whole.

While we were adjusting sampling weights, it was a simple matter to reduce the potential bias from omitting small firms. For example, any sample that excludes firms with annual sales under \$10 million will underweight retailing. We simply oversampled the remaining retail firms enough to assign retailing its proper national weight—and similarly for every other two-digit industry. In this way, we eliminate from our sample the portion of any “large-firm bias” that stems from the different industrial structures of small versus large firms. However, to the extent that small firms really differ from large firms *in the same industry*, the bias remains.

IV. Preliminary Results

The interviewing is now in progress. Given the limited clerical and interviewer resources at my disposal, we are conducting them in eight waves of approximately 50 letters (25 interviews) each over a period of about 18 months. As of this writing, the first three waves are nearly complete and we have just begun the fourth.

The results reported here are based on tabulations of the 72 interviews completed as of mid-November 1990. Since this partial sample is too small to permit much useful disaggregation, I restrict my attention to national averages here. Once the full sample of 200 is available, however, there will be a great deal of information about, for example, how the validity of different theories varies by industry.

The questionnaire comes in two parts and usually takes about 45–60 minutes to administer. The first part gathers a variety of factual information about each firm, while the second inquires directly about the theories. The factual data will eventually be

TABLE 1—MEASURES OF PRICE STICKINESS^a

A. Frequency of Price Change ^b		Percent of Companies
More than 12		10.1
4.01 to 12		4.3
2.01 to 4		10.1
1.01 to 2		20.3
1.0		37.7
Less than 1.0		17.4
B. Question: How much time normally elapses after a significant _____ before you raise (reduce) your prices?		
Event	Mean Lag ^c	Standard Deviation
Increase in demand	3.23	2.93
Decrease in demand	3.60	3.93
Increase in costs	3.17	2.90
Decrease in costs	3.97	4.47

^aBased on 72 interviews completed as of mid-November 1990.

^bTimes per year.

^cIn months.

used for disaggregation, in cross tabulations, and as right-hand variables in regressions explaining the validity of the various theories. I make little use of these data in this paper, but one issue is important enough to deal with right now: Just how “sticky” are prices in the U.S. economy?

To find out, we ask each firm, “How often do the prices of your most important products change in a typical year?” The distribution is shown in panel A of Table 1. These results offer cheer for Keynesians who model prices as, say, rigid for one period or more. I like to joke that in macroeconomic theory we know that the length of “the period” is one quarter. According to these data, less than 15 percent of the GNP is repriced more frequently than quarterly, and fully 55 percent is repriced no more often than once per year. If we had to pick a single abstraction to represent the whole economy, annual price change (the median response) would appear to be the “right” model.

From the point of view of macroeconomic theory, frequency of price change may not be the right question to ask, for it depends as much on the frequency of shocks as on

firms' pricing strategies. We are more interested to know how long price adjustments lag behind shocks to demand and cost. Panel B summarizes the results from a series of questions of the form, "How much time normally elapses after a significant _____ before you raise [reduce] your prices?" We inquire about four events: an increase in demand, a decrease in demand, an increase in cost, and a decrease in cost. The table shows that, while the distributions are quite spread out, the mean lags cluster in the 3–4 month range. There is precious little evidence that prices increase faster than they decrease, and virtually none that firms respond to cost shocks more quickly than to demand shocks.

The bulk of the questionnaire inquires about the twelve theories themselves. We begin each section by succinctly summarizing one of the theories and then posing a question like, "How important is this in explaining the speed of price adjustment in your company?"⁴ Respondents answer freely in their own words, but interviewers code the responses on the four-point scale indicated at the bottom of Table 2.

While the scale resembles the typical grading system used at universities, we should expect the "grades" to be more compressed. At the high end, no theory is ever going to score a perfect 4.0. A theory rated "very important" by half the firms and "of minor importance" by the other half would be a fine theory indeed. So I would interpret a score of 3.0 as excellent. At the low end, a score of 1.0 would mean that every firm in the sample dismissed the theory out of hand. This is not a grade of D, but a kind of super F. Indeed, any score below 1.5 is a remarkably poor performance.

If the respondent says a particular idea is "totally unimportant" to the speed of price adjustment, we ask no further questions about that theory. Otherwise, we ask a variety of follow-up questions (sometimes many,

TABLE 2—SUMMARY EVALUATION OF THEORIES^a

Rank	Theory	Mean Score ^b	Percentage "3" or Higher
1	Delivery lags/ service	2.86	65
2	Coordination failure	2.85	66
3	Cost-based pricing	2.72	57
4	Implicit contracts	2.52	56
5	Explicit nominal contracts	2.29	40
6	Costs of price adjustment	2.28	43
7	Procyclical elasticity	1.97	36
8	Pricing points	1.97	33
9	Inventories	1.72	28
10	Constant marginal cost	1.56	19
11	Hierarchies	1.54	18
12	Judging quality by price	1.45	12

^aBased on 72 interviews completed as of mid-November 1990.

^bBased on the following scale: 1 = totally unimportant; 2 = of minor importance; 3 = moderately important; 4 = very important; N = don't know or cannot answer.

sometimes few) geared specifically to each theory. In this interim report, I ignore follow-up questions and concentrate on the main questions about the perceived validity of each theory.

Table 2 summarizes the central results on the validity of the twelve theories, as perceived by actual decision makers, ranked in order of popularity. The second column names the theories briefly, and sometimes cryptically. (Explanations follow.) The third column records the mean score on the four-point scale; remember that we expect most theories to score in the 1.5–3.0 range. We can interpret a rating of 1 or 2 as meaning that the firm rejects the theory as an explanation of price rigidity, and a rating of 3 or 4 as meaning that the firm accepts it. So the fourth column gives the percentage of respondents that rate the theory as 3 or higher—an indicator of the fraction of the private, for-profit GNP to which this theory

⁴For many theories, there is a preliminary factual question. For example, before we ask whether judging quality by price deters price increases, we first ask whether the firm thinks its customers actually do judge quality by price.

applies. The two alternative ways of ranking the theories agree closely. Ranked from best to worst by mean score, the theories are:

1) *Delivery lags / service*:⁵ The idea here is that price is but one of several elements that matter to buyers. Rather than cut (raise) prices when demand is low (high), firms might prefer to shorten (lengthen) delivery lags or provide more (less) auxiliary services. The mean score for this theory so far is a healthy 2.86. Seventy-six percent of the firms accept the premise, and 65 percent say it is an important factor in slowing down price adjustments.

2) *Coordination failure*: This very old idea has been revived and formalized in recent New Keynesian theorizing. The notion is that firms might like to raise or lower prices, but hesitate to do so unless and until other firms move first. Once other firms move, they follow quickly. The mean score for this theory of 2.85 and its "acceptance rate" of 66 percent are quite high indeed.

3) *Cost-based pricing* connotes another old Keynesian idea: that prices are based on costs and do not rise until costs rise. Well over half the firms in our sample so far rate this a moderately or very important factor in explaining the speed of price adjustment; the mean score is 2.72.

4) *Implicit contracts* connotes the "invisible handshake" theory that firms have implicit understandings with their customers which proscribe price increases when markets are tight. Though this theory obtains an average score of 2.52, it elicits strikingly bimodal responses. Sixty-one percent of the sample accepts the premise that implicit contracts exist and, within this subsample, respondents generally think such contracts are an important source of price stickiness: the mean response within this group is a stunning 3.42.

5) *Explicit nominal contracts* refers to the naive Keynesian idea that written contracts prohibit price adjustments while they remain in force. Most firms have such con-

tracts, for at least some of their products; but discounting appears to be common. Our rough estimate is that this is an important factor in price stickiness for only about 40 percent of the economy. The mean score (counting firms with no explicit contracts as 1's) is 2.29.

6) A number of theorists have suggested that firms incur special *costs of price adjustment* whenever they change prices. Both menu costs and convex costs are lumped together under this rubric, but follow-up questions distinguish between the two. About 70 percent of all firms report that they have such costs. But fewer than half think them an important factor in slowing down price responses. The average score is 2.28.

7) According to a very old idea revived in the mid-1980's, demand curves become less elastic when they shift in. Such *procyclical behavior of elasticity* would lead to countercyclical markups which would, in turn, rigidify prices. This is another case where the premise is often accepted (59 percent of the time), but the idea is thought to be a significant source of price rigidity in only 36 percent of the cases. The average score is just below 2, which connotes minor importance.

8) *Pricing points* refers to the idea that certain prices (such as \$19.95 for a shirt) are psychological barriers that firms are reluctant to cross. This also scores 1.97 on average. Again, a majority of the firms (53 percent) accepts the premise; but only a third think it explains much price stickiness.

9) *Inventories* refers to the theory that, when demand rises (falls), firms draw down (build up) their inventories rather than increase (decrease) prices. Note that this question is not asked of service companies, so the sample size so far is only 33. Of these, however, fewer than a third think inventories a significant factor in deterring price adjustments. The average score is a low 1.72.

10) *Constant marginal cost* is shorthand for the notion that prices are sticky because both marginal costs and markups are constant over the business cycle. Since its ratings are so poor (a mean score of only 1.56

⁵Space constraints prohibit me from giving scholarly references for each of the theories. I apologize to those slighted.

and an acceptance rate of just 19 percent), I should explain how we posed the question. First, we asked respondents how their "variable costs of producing additional units" (our plain-English translation of marginal cost) behave as output rises. Only 42 percent of the sample reported that their MC was constant. Among this minority, about 60 percent said that the constancy of MC was totally unimportant or of minor importance in explaining their speed of price adjustment.

11) *Hierarchies* is a code word for a "theory" that does not come from the academic literature. Rather, it was suggested to us by an executive of a large corporation. The simple idea is that price changes are slowed down by the difficulty of getting a large, hierarchical organization to take action. Its low average score of 1.54 may appear unsurprising, since it should apply only to giant companies. But, even among firms with annual sales over \$100 million, its mean score is only 1.66.

12) The worst theory, according to our practitioners, is *judging quality by price*, a theory I have mentioned before. It obtains a paltry mean score of 1.45 and an acceptance rate of just 12 percent. Less than a quarter of the firms in the sample so far think their customers would "interpret a price cut as a signal that the quality of the product has been reduced." In our tabulations, we count the others as rating the theory "totally unimportant." However, among the minority of firms whose customers *do* judge quality by price, roughly 45 percent say it is a moderately or very important factor in dis-

couraging price increases. Thus this theory may be of some importance within a narrow sector of the economy.

In summary, it seems to me that the theories divide themselves into three groups. The top four listed above distinguish themselves as especially promising and/or applicable to the U.S. economy. The leading theory (by a slim margin), delivery lags/service, has not received the attention it deserves. The other three look to me decidedly Keynesian. Cost-based pricing is, of course, old-fashioned Keynesian stuff. Coordination failure is a major strain of New Keynesian theorizing. And the "invisible handshake" is by now part of the Keynesian tradition.

At the bottom of the list, we find four theories that appear to be rejected by our respondents. Two of them have enjoyed great popularity in the 1980's. If practitioners are to be believed, marginal costs are constant in only a minority of industries. And judging quality by price appears to be neither a common nor an important phenomenon.

It is true that these results are based on only 36 percent of the ultimate sample. But I will be surprised if the rankings change dramatically as more data come in; 72 observations is sufficient to answer most questions about national averages. The payoff to a large sample will come later, when I have enough data to answer disaggregated questions like, "What kinds of firms make 'invisible handshakes' with their customers?" I will report information of that sort at a later date.

DISCUSSION

ROBERT J. SHILLER, Yale University: This is exciting research. Theories of price stickiness are fundamental to much of macroeconomics. Surely it is a good idea to let those who set prices explain why they do what they do. These people tend to be professionals who have reason to think carefully about their reasons. Alan Blinder has done a very creative job of finding out what they can tell us. I hope that this is the beginning of many studies that use similar methods.

Blinder is working here to rectify the damage that overliteral interpretation of Milton Friedman's theory of positive economics has wrought in the economics profession. Many people seem to have thought that Friedman's "billiard player" analogy justifies omitting ever asking people about what they do. Friedman may be right that one cannot ask the player to explain why some shots are effective, and that a theoretical physicist could explain better. But, on the whole, I think that it would be a disastrous mistake to ask a physicist to model the behavior of a billiard player without allowing the physicist to get the player's help in the modeling process. The physicist will not understand the strategy of the game, will not know what the player's short-run objective would be on a given shot, and will likely omit considerations such as english or margins of error that may be difficult to theorize about or about which the physicist does not have full information.

The whole premise of the billiard player analogy that we economists should assume that people know how to optimize is also wrong. In many of the important economic decisions people make, such as the decisions on pricing strategy discussed here, a real evaluation of the outcome is not possible for many years, and even then the long-run effects cannot be disentangled from many other factors that are changing through time. Lacking such evaluations, people must satisfy or imitate successful people. There are hunches or popular models of the world that guide people, and we had best find out what these are:

Instead of "positive economics" as that term is commonly interpreted, we should instead follow what Edward Leamer calls "Sherlock Holmes inference," looking at every scrap of evidence and being alert to the evidence that even "small" or "unreliable" clues provide. Fortunately many economists seem to heed Leamer's advice, but the evidence obtained by asking people about the reasons for their behavior is not commonly used.

The research presented here should provide a good launching pad for further work in the future. As a pioneering study, Blinder's paper takes a broad view; much refinement should follow.

I was struck that the number one theory (the theory that delivery lags and service vary instead of price) may not be so much a theory of price rigidity of interest to macroeconomic theorists as it is a theory that the prices we *measure* are more rigid than true prices, particularly if we consider the service cuts part of this theory. If, instead of raising a price, a firm imposes costs on the buyer in the form of declining service, so that the buyer must pay for this service somewhere else, then the "true" price has changed, no "excess demand" should be generated by the failure to raise the nominal price. Thus, further research might ask decision makers directly why they sometimes allow a situation of excess supply or demand to persist, rather than asking about price rigidity *per se*.

We could also adjust our research strategy to try to understand why the standard business practices survive as they do, through time and against the occasional testing by innovators. To learn about this, it would be good to ask price setters about actual experiments that were made in *changing* price-setting behavior. Ask them to remember any story about someone who tried arguing for a different policy. Track that person down, and get the story as to what lessons were learned from the experience. Doing this is perhaps much better than asking the current price setters themselves, who might only report the vaguest

recollections of the outcome of past experiments.

Further research might also clarify the importance of fairness considerations in price-setting behavior. About half of Blinder's twelve theories might be interpreted as relating to concerns about possibly unfair behavior of price changers; indeed a number of questionnaire items concern whether they think the behavior would "antagonize" customers. Further attention could be focused on these questions. It seems likely that some rigidity of retail prices is due to such attitudes on the part of consumers.

Subsequent studies along lines initiated by Blinder might each focus in on one of the twelve theories, and spend more time eliciting information from respondents about the relevance of these theories. Respondents cannot give their full information about any of these theories without time to think and without substantial prodding. Much could also be learned by comparing answers to questions such as these across countries or through time. I think that it would be a good idea to supplement this endeavor with a similar survey of individual consumers. Their behavior is as important in many of these stickiness questions as that of price setters. There is a lot more work to do along lines suggested by Blinder's study, but this promises to be an excellent start.

ROBERT J. GORDON, Northwestern University: Alan Blinder's project is refreshingly novel. Instead of proposing yet another theory of price stickiness, he suggests that interviews with business people will prove the decisive filter to sort the theories. Although my own research predisposes me to sustained applause, I am happier with the point of departure of Blinder's project than with its detailed execution. To be fair, the critic cannot just append a list of additional questions that Blinder should have asked the respondents. As economists, we must treat interviewing time as a scarce resource. With the same 45 minutes or an hour, what would we have done differently?

My objections are fourfold. First, I question the basic approach of running a beauty

contest among theories, because there are too many overlaps among the poorly defined alternatives that are thrown at the respondents. For instance, a typical firm may adjust service and delivery lags, experience coordination failure, and base prices on costs, indicating that the most popular "theories" are complementary, not mutually exclusive. Second, Blinder exaggerates the degree to which the competing theories are "empty", and misses a unique opportunity to distinguish among theories by probing deeper. Third, many of the questions are vulnerable to the failure to specify time dynamics carefully. Fourth, although the project is spaced out over two years, there is no "funnel" principle by which early results lead bad theories to be discarded, so that the released time is used to probe more deeply into the good theories.

Here's an alternative approach that would have yielded more out of the same interviews. For price stickiness to matter for business cycles, we must observe a failure of the aggregate price level to mimic fluctuations in nominal GNP. To use this notion to sort theories, we can introduce the four-way distinction among aggregate and local demand and supply shocks. Then we would ask respondents how much they would change price over a specified period after a specified event involving one or more of the four shocks.

For instance, we might ask hypothetically what happens to price if nominal GNP drops by 5 percent but the firm's revenue and costs do not change. If the respondents answer that no price change is warranted, then we have important evidence to bolster the old micro truism that firms care about the relation of price to cost, not price to nominal GNP. Then we might ask, when nominal GNP drops by 5 percent, why don't you assume that your costs will drop by 5 percent? If not now, over what horizon? This would tell us something about the underpinnings of coordination failure, and what I call the input-output table approach. We then might ask hypothetically what happens to price if aggregate and local nominal demand rise by 5 percent, while costs are constant, versus an opposite situation in

which costs rise by 5 percent while local and aggregate demand remain constant.

If the respondents say in the fashion predicted by Okun that they raise price more in response to costs than demand, we could then ask them about menu costs. I would ask every firm why they change prices infrequently, while the price of avocados and navel oranges can change daily. What's the crucial difference between them and the produce department of the local supermarket? One could ask each firm why it changes price so infrequently—is it that menu costs are so high, or that demand and/or costs change at such a steady and predictable rate? How much would demand and/or costs have to jump to cause the firm to break out of the annual pattern and move to shorter intervals of price changes (we know, after all, that prices change much faster in Brazil than in the United States). Throughout this investigation, one would want to dig beneath the trite but true explanations 2 and 3 in the Blinder ranking (coordination failure and cost-based pricing) to ask why firms fail to coordinate, why they do not change price to mimic nominal GNP, why they fail to assume costs will not mimic nominal GNP, and whether they react instantly to every cost change or lump the reaction to cost changes into a single infrequent price change, and why.

Since the emphasis of Blinder's paper is on the virtues of interviews and on survey design, it is helpful to recall the most famous antecedent of his project, the famous 1939 study by Hall and Hitch. Their classic paper is a good model to suggest changes in Blinder's approach. Instead of conducting a popularity poll, they start with a single theoretical point of departure for the questions, the standard theory that marginal cost should equal marginal revenue, and then, building on the discovery that full rather than marginal cost matters, ask why so? They come to a profound conclusion that has been the basis for much Keynesian thinking since 1939, "prices... will be changed if there is a significant change in wage or raw material costs, but not in response to moderate or temporary shifts in demand" (p. 33). Is this still so, and if so, why?

HERSCHEL I. GROSSMAN, Brown University: Properly designed interviews can be valuable sources of information about how people behave. But, the ability of interviews to reveal why people behave as they do (or, more precisely, to discriminate among different models of decision making) is more problematic. Alan Blinder's defense of his interview survey against possible methodological criticisms does not recognize this distinction between the gathering of facts about behavior and the explaining of facts about behavior. This failure reflects what seems to me to be Blinder's underlying ambiguity about his objectives and about what he means by a "theory" of price stickiness.

When we speak of a theory about economic behavior, we usually have in mind a connected set of assumptions and derived implications about the decision-making process of economic agents. In neoclassical economics, these assumptions involve a specification of preferences and constraints and the derivation of implications involves the solution of a constrained-maximization problem. As we can interpret any behavior that we observe to be the solution to some constrained-maximization problem, a central part of the research program of neoclassical economics that we as neoclassical economists are trained to carry out is an attempt to discover the particular constrained-maximization problem that best fits the facts. Because the facts with which economists are concerned typically involve observations about markets or economies, rather than about the idiosyncratic behavior of specific individuals, we are usually satisfied if a theory fits the facts in an "as-if," rather than literal, sense.

With a couple of possible exceptions, the twelve "theories" of price stickiness that Blinder sets out are not theories in the sense of models of decision making. Blinder's theories are mainly hypotheses about how people behave rather than about why they behave as they do. Accordingly, most of the information that his interviews elicit does not directly address the neoclassical question: What specifications of preferences and constraints are likely to be useful for building constrained-maximization

models of price setting that are empirically relevant?

One possible exception to the criticism that Blinder's theories are not attempts to explain behavior is the theory that he denotes "coordination failure." Theorists have formulated coherent models in which difficulty in coordinating strategies constrains price changes. But, the questions that Blinder poses to his interviewees do not seem to get to the factors that distinguish such theories. Specifically, the unsurprising fact that most firms report that "they are afraid to get out of line with what they *expect* competitors to charge" (emphasis added) does not by itself imply that coordination problems actually impede price adjustments.

The other possible exception to my criticism of Blinder's theories is the theory that he denotes "cost of price adjustment." It is interesting to know that "fewer than half" of the interviewees report that costs of price adjustment are an important constraint on pricing decisions. But, even so, such a constraint might still be a useful component of a model that is true in the as-if sense.

Blinder's other theories, especially the ones he denotes "delivery lags/service," which receives the highest score from his interviews, "cost-based pricing," as well as the ones he denotes "implicit contracts" and "explicit nominal contracts," describe the results of interactions between buyers

and sellers, but do not address why buyers and sellers behave as they do. Delivery lags, cost-based pricing, and contractual arrangements, whether implicit or explicit, are symptomatic of underlying preferences and constraints. For example, delivery lags and/or contracts might be symptoms of coordination failure or of adjustment costs. Theories that attempt to explain behavior would specify the preferences and constraints that lead to delivery lags and cost-based pricing, or that produce the choices that are codified in contracts.

One final constructive criticism: I strongly endorse Blinder's view that "Keynesian economics is...the economics of nominal rigidities." The distinctive emphasis on the stickiness of nominal wages and prices and on the associated possibility that monetary disturbances can cause markets to fail to clear explains why Keynesian economics remains central to the modeling of macroeconomic phenomena. For this reason, I would have liked Blinder's survey to have included questions that would have distinguished between nominal and real rigidities.

In sum, Blinder's interviews provide information about pricing behavior, but little or no information about the reasons for this behavior. Is the problem that interviews are not capable of addressing the questions to which we really want answers? Or, has Blinder not yet fully exploited the interview methodology?

Projecting Faculty Retirement: Factors Influencing Individual Decisions

By G. GREGORY LOZIER AND MICHAEL J. DOORIS*

The impending elimination of mandatory retirement for tenured faculty on January 1, 1994 (Public Law 99-592, 1986), has raised the visibility of a number of questions about faculty retirement behavior. What are the factors that influence individual faculty members' retirement decisions? How important are financial and nonfinancial considerations? Why do faculty in private institutions work to a later age, on average, than their colleagues in public institutions? Are there other systematic (for example, gender- or discipline-related) differences as well?

This paper is an attempt to answer such questions about individual faculty retirement behavior. The paper utilizes data collected as part of a comprehensive national study that projected faculty retirements through the year 2003 for over 35,000 faculty at 101 doctoral research, comprehensive, and general baccalaureate institutions (see our 1990 paper). The discussion that follows uses data from that broader institutional survey and from a survey of 747 faculty members age 55 and over who had separated from this same set of 101 institutions. Among the information provided by the 518 usable responses were data on factors that influence faculty members' decisions regarding the appropriate time to retire.

I. Factors Influencing Retirement Decisions

Previous Studies. Some higher education studies have touched on individual faculty

retirement behavior, but much of what is known is drawn from research in other employment sectors. For example, an empirical study by H. Anderson et al. (1986) determined that the retirement decision among the general workforce is influenced by both the current value of relevant variables such as income, inflation, and health, and by expectations about their future value. They found that retirement decisions within the general work force are most likely to be altered by changes or anticipated changes in such factors as Social Security benefits, recession or inflation, unemployment rates, and rates of economic growth. This finding reinforces the results of an earlier report by S. F. Quinn and R. V. Burkhauser (1983) in which they found that financial considerations such as Social Security benefits were critical factors influencing retirement decisions. Similarly, S. LaRock (1987) noted that, in addition to such determinants as health, workplace interaction, and intensity of family ties, perceived adequacy of retirement income and the financial status and retirement plans of one's spouse were the most critical elements influencing the retirement decision.

In the higher education sector, among the major findings of three studies conducted for the Consortium on Financing Higher Education (COFHE) was the observation that the most powerful predictor of delayed retirement was "fear of inadequate income during the first two years of retirement" (S. Montgomery, 1989, p. 57). The most frequently cited reason for postponing retirement, as reported in one of the studies, was the need for continued full-time salary.

Study Results. The present study used 18 factors of potential influence on the retirement decisions of faculty members. These factors were derived from those iden-

[†]*Discussants:* W. Lee Hansen, University of Wisconsin-Madison; Olivian Mitchell, Cornell University.

*Executive director and senior planning analyst, respectively, Office of Planning and Analysis, Pennsylvania State University, University Park, PA 16802.

tified in the studies cited above as being important to the general workforce. In the survey, respondents indicated on a Likert scale the extent to which each of the 18 factors was not important (1) or very important (6) in their decisions to retire. The factors and their average ratings (in parentheses) were:

1) Overall financial status (4.4); 2) Eligibility for full retirement benefits (4.4); 3) Desirability of more personal/family time (4.1); 4) Other interests (3.5); 5) Working conditions and policies (2.9); 6) Availability of early retirement incentive benefits (2.6); 7) Personal health (2.5); 8) Annual salary increases (2.2); 9) Availability of emeritus benefits (2.2); 10) Mandatory retirement policies (2.1); 11) Other employment opportunities, including self-employment (2.0); 12) Health of a spouse (1.9); 13) Administrative pressure (1.8); 14) State of the economy (1.8); 15) Interaction with coworkers (1.7); 16) Budget cutbacks (1.7); 17) Curricular revision (1.6); 18) Timing of spouse's retirement (1.5).

Every factor received ratings at both extremes of the scale. In other words, each was important to some respondents and not important to others. Even so, the two most salient factors emerging from the responses were overall financial status and eligibility for full retirement benefits. In addition, the availability of early retirement incentive benefits, while not an important variable (and perhaps unavailable) to 60 percent of the respondents, was the most important factor in the retirement decision for over 20 percent of the faculty retirees. When such programs are available, they become a powerful variable in combination with other financial considerations.

Some distinctions were identified when the ratings were analyzed by discipline and gender. Statistically significant relationships between the retirement factors and discipline were found for four factors: 1) desirability of more personal/family time; 2) working conditions and policies; 3) state of the economy; and 4) budget cutbacks. Education faculty were more likely to see the need for more personal/family time and the state of the economy as more important factors in the retirement decision

than faculty in other disciplines. Agriculture faculty were more likely to make the retirement decision based on working conditions and policies and budget cutbacks. Library sciences faculty were also more likely to consider working conditions and policies in their retirement decisions. Although the "other employment opportunities" factor did not produce an overall significant difference by discipline, it was interesting to note that all library sciences faculty retirees responded to this item with a minimum score of 1.0, that is, not important.

Separate analyses of the retirement decision factors by gender also identified four statistically significant variables: 1) other interests; 2) working conditions and policies; 3) administrative pressure; and 4) interaction with coworkers. The greatest variability between male and female responses on these four factors was on working conditions and policies (2.8 vs. 3.4). The average rating for "other interests" for both females and males was 3.5. However, the male responses clustered around the average, while those of the females tended to be at one extreme of the ratings scale or the other.

It is important to reiterate that, even when there were statistically significant differences by either discipline or gender, financial considerations were still the predominant factors in the retirement decision.

II. Impact of Retirement Program Type on Retirements

One determinant of "overall financial status" in retirement is the income replacement value provided by an individual's retirement plan annuity. Most retirement plans in higher education are one of two types: defined-benefit or defined-contribution programs. Defined-benefit plans are typical of state employee retirement systems while defined-contribution plans are associated more with private retirement programs, the best known of which is TIAA-CREF. According to F. P. King and T. J. Cook:

[T]he *defined benefit* approach fixes benefits in advance as a percentage of salary for each year of service or, as in

some industrial plans, a flat dollar amount per month or year of service. The *defined contribution* approach fixes contributions in advance as a percentage of each employee's salary, allocates these contributions to individual accounts, and pays benefits on the funds credited to each participant.

[1980, p. 44]

In short, the former fixes benefits, while the latter fixes contributions.

A number of studies on faculty retirement trends have established a relationship between institutional control, that is, public vs. independent, and average age of retirement (Consortium on Financing Higher Education, 1987; W. L. Hansen and K. C. Holden, 1981; our 1988-89 article and 1990 paper). These studies have shown that the faculty retirees in independent institutions are, on average, two to three years older than their peers in public colleges and universities. Although multiple factors likely contribute to this difference, there has been some speculation that one of these may be type of retirement program.

How likely is it that the public-private retirement age difference is related to the defined benefit-deferred contribution distinction? The evidence is mixed. Holden and Hansen, noting the difference in retirement trends between public and independent institutions, have developed cost analyses to assess whether type of retirement program may be one of the multiple factors influencing differences among institution types. In their model they observed that "despite very different plan structures, annual benefits from defined-benefit and defined-contribution plans increase comparably, on the average, as retirement is postponed" (1989, p. 79). They concluded that the decision to defer retirement is not likely to be caused by enrollment in one type of plan.

A recent TIAA-CREF *Research Dialogues* (1990) explored the issue further using a measure called the "salary replacement ratio." This is simply the ratio of retirement income to preretirement salary. Examining two different scenarios, one in which the individual enters the retirement plan at age 30 and another in which the

individual enters the retirement plan at age 45, the analysis looks at the salary replacement ratio for retirement at ages 65 through 70, for both defined-contribution and defined-benefit plans. The TIAA-CREF analysis reveals "that a delay of retirement beyond normal-age 65 can be expected to provide somewhat higher initial benefits in subsequent starting years under both the defined contribution and defined benefit plans" (p. 5). However, the average replacement ratio under the defined-benefit plan is approximately 9 percent lower than the defined-contribution program.

Retirement Plan Retirement Age Differences. The present study did not attempt to model actual retirement benefits derived from defined-benefit and defined-contribution programs. The study was set up to collect data on age at retirement by institutional control and type of retirement program. In particular, the institutional survey segment of this study (101 responding institutions) was geared in part to the question of whether the type of retirement program had any relationship to average age of retirement. In order to separate the institutional control effect on type of retirement program, separate average retirement age calculations were developed for all retirees in each of four clusters:

Public defined-benefit retirees, 63.1 ($n = 475$); public defined-contribution retirees, 65.4 ($n = 580$); independent defined-benefit retirees, 65.4 ($n = 6$); independent defined-contribution retirees, 65.3 ($n = 1241$).

The average retirement age among public institution retirees is over two years lower, 63.1 vs. 65.4, for defined-benefit retirees. The 65.4 average retirement age for public defined-contribution retirees is also extremely close to the higher 65.3 average age for independent defined-contribution retirees. (Because so few independent institutions offer defined-benefit programs, there were only 6 independent defined-benefit retirees, a number too small from which to draw any conclusions.)

Explaining the Difference. Although the data presented above do not establish a direct cause and effect relationship, the results suggest some interactive effect be-

tween type of retirement program and the retirement decision. However, it is not clear that in all cases early retirement is more likely under one type of plan than the other.

Both the Holden-Hansen and TIAA-CREF analyses discussed above point out that defined-contribution and defined-benefit plans are designed to produce fairly comparable levels of retirement income for individuals with similar years of service and income history. Yet, the TIAA-CREF discussion also points out that at some point future retirement income increases more rapidly under the defined-contribution program. "Return on investment" is a useful concept for examining further whether under this circumstance one type of plan is more likely to encourage earlier or later retirement.

The *investment* would be the additional years worked beyond "normal" retirement age, while the *return* would be measured as the additional retirement income. In the case described in the TIAA-CREF *Research Dialogues*, retirement income for working beyond age 65 increases more rapidly under the defined-contribution program than under the defined-benefit plan. Assuming, on the one hand, that the prospective retiree has established a desirable retirement income level necessary to trigger the decision to retire, the defined-contribution participant could achieve this level sooner than a defined-benefit participant. Under this scenario, the defined-contribution plan would provide a good return on the investment of additional years of service and would encourage an earlier retirement, as the prospective retiree meets the preset retirement income goal sooner than would have been realized under the comparable defined-benefit plan.

On the other hand, the slower post-65 replacement ratio growth rate under the defined-benefit plan could encourage the participant to retire earlier *if* the employee determined that the amount of increase was not sufficient to warrant continued employment, that is, the marginal return was not worth the investment of each additional year worked. The latter factor could help to explain why participants in some defined-benefit plans have suggested that they re-

tired when they did because they would have "continued to work for free." To the extent that defined-benefit programs have provisions that eliminate penalties for retiring after 30 or 35 years of service, or have other provisions that limit or reduce the rate of growth in anticipated retirement income, this latter scenario may contribute to the explanation for the younger retirement age for faculty members on defined-benefit annuities.

III. Discussion

In the face of the impending elimination of mandatory retirement in 1994 for all colleges and universities that have not already done so, institutions have shown considerable interest in devising mechanisms for influencing retirement rates. Because of the importance of individual financial status upon the retirement decision, financial inducements are probably the most powerful tool for influencing that decision. In fact, a financial response in the form of incentive early retirement programs was the initial reaction of many institutions in the late 1970's and 1980's as the permissible mandatory retirement age shifted from 65 to 70 (J. L. Chronister and T. R. Kepple, 1987). Whether incentive early retirement programs are the correct response in light of more recent information regarding faculty age distributions and anticipated retirement rates has been the subject of debate. It is questionable whether incentive programs that become permanent or are available to all prospective retirees produce the desired effects. There is speculation that the more useful programs have been those that have been limited in both duration and scope. Institutions may wish to analyze correlations between retirement income and retirement age to determine the likely effect of incentive programs or other measures.

Beyond confirming the importance of financial considerations, this study suggests that virtually all factors that bear upon the retirement decision are relevant to some faculty but not to others. Retirement is a very personal decision. Variables such as the faculty member's health or the health of a spouse, other professional consulting and

employment opportunities, and the need for more personal time cannot be controlled by the institution. Similarly, budget cutbacks at the state or federal level, inflation and recession, or changes in Social Security provisions are outside the realm of institutional policy.

But other factors, including several that differ significantly in importance among faculty on the basis of discipline and/or gender, can be influenced by the institution. For example, working conditions and policies make a difference in the retirement decisions of library science faculty and of women. Women also noted that administrative pressure and interaction with coworkers affect the desirability of continued employment.

Finally, the results of this study clearly showed that, perhaps as expected, money, as a factor in the retirement decision, matters to nearly everyone. The effects of less tangible elements of professional satisfaction are not as uniform or consistent, but nonetheless they matter also. Career and retirement decisions are ultimately individual decisions, and the significance of personal interactions, the work environment, leadership, and so on cannot be discounted. The extent to which deans and department heads can recognize and act upon the full range of financial and nonfinancial variables is of critical importance in influencing personal career and retirement behaviors of faculty members.

REFERENCES

- Anderson, H., Burkhauser, R. V. and Quinn, J. F., "Do Retirement Dreams Come True? The Effect of Unanticipated Events on Retirement Plans," *Industrial and Labor Relations Review*, July 1986, 39, 518-26.
- Burkhauser, R. V. and Quinn, J. F., "Is Mandatory Retirement Overrated? Evidence from the 1970s," *Journal of Human Resources*, Summer 1983, 18, 337-59.
- Chronister, J. L. and Kepple, T. R., *Incentive Early Retirement Programs for Faculty*, ASHE-ERIC Higher Education Report No. 1, Washington, 1987.
- Hansen, W. L. and Holden, K. C., "Mandatory Retirement in Higher Education," unpublished report for the U.S. Department of Labor, University of Wisconsin-Madison, 1981.
- Holden, K. C. and Hansen, W. L., *The End of Mandatory Retirement: Effects on Higher Education*, New Directions for Higher Education, No. 65, Spring 1989.
- King, F. P. and Cook, T. J., *Benefit Plans in Higher Education*, New York: Columbia University Press, 1980.
- LaRock, S., "Retirement Patterns 1978-86: ADEA Protection Has Little Effect on Employee Decision," *Employee Benefits Plan Review*, August 1987, 42, 30-34.
- Lozier, G. G. and Dooris, M. J., "Elimination of Mandatory Retirement: Anticipating Faculty Response," *Planning for Higher Education*, No. 2, 1988-89, 17, 1-13.
- _____ and _____, "Faculty Retirement Projections Beyond 1994: Effects of Policy on Individual Choice," paper presented at the American Association for Higher Education national conference, San Francisco, April 1990.
- Montgomery, S., "Findings from the COFHE Studies," in K. C. Holden and W. L. Hansen, eds., *The End of Mandatory Retirement: Effects on Higher Education*, New Directions for Higher Education, No. 65, Spring 1989, 51-62.
- Quinn, S. F. and Burkhauser, R. V., "Influencing Retirement Behavior: A Key Issue for Social Security," *Journal of Policy Analysis and Management*, No. 1, 1983, 3, 1-13.
- Consortium on Financing Higher Education (COFHE), "Early Retirement Programs for Faculty: A Survey of Thirty-Six Institutions," Cambridge, MA, 1987.
- TIAA-CREF, "Income and Other Factors in Delayed Retirement," *Research Dialogues*, No. 24, New York: Teachers Insurance and Annuity Association-College Retirement Equities Fund, January 1990.
- Public Law 99-592, "Age Discrimination in Employment Act Amendments of 1986," 31 October 1986.

Ending Mandatory Retirement in the Arts and Sciences

By SHARON P. SMITH*

No institution interested in preserving quality can tolerate a growing gerontocracy that necessarily brings with it declining productivity. The disastrous effect on young scholars surely needs no elaboration. If ever mandatory university retirement is deemed to be age discrimination, an alternative mechanism will have to be found to accomplish the same purpose.

[Henry Rosovsky, 1990, p. 211]

On January 1, 1994, the elimination of mandatory retirement for tenured faculty (uncapping) which was enacted in the 1986 amendments to the Age Discrimination in Employment Act (ADEA) will take effect. This grace period between enactment and effective dates was a response to the concerns that have been regularly expressed by educational administrators such as Rosovsky. In the congressional debate concerning the impact of uncapping on the retirement behavior of tenured faculty and the implications of any behavioral changes for the vitality of higher education, it was concluded that this delay would allow time for study, adjustment, and the opportunity to request a permanent exemption, if necessary.

The concerns voiced by educational administrators have several dimensions. In the absence of a mandatory retirement clause, academic tenure becomes a lifetime employment contract. Under these conditions, the central concern is whether in a world in which all retirement is voluntary, no faculty

will volunteer. This anxiety raises concerns over the implications of such a change in behavior for: job opportunities for new faculty;¹ the success of affirmative action initiatives that call for the integration of women and minorities into a faculty population that was historically largely white male; the cost of predominately senior faculty in which each member is presumed to be relatively highly paid; and the quality of education provided by faculty whose abilities and whose timeliness of ideas may be diminishing with age.

The Project on Faculty Retirement (PFR) was established in response to these concerns.² It was sponsored by the American Association of University Professors, the Association of American Universities, the Consortium on Financing Higher Education, and the National Association of State Universities and Land Grant Colleges to analyze the implications of uncapping for tenured faculty in the arts and sciences.³

The PFR's research had three thrusts: an analysis of the current and estimated future retirement behavior of tenured faculty; an inquiry into the attitudes towards retirement among tenured faculty who have already retired and those who are nearing the age for a retirement decision (55 and over); and a study of the relationship between chronological age and a variety of indicators of faculty vitality (student teaching evalua-

¹With a growing consensus that higher education faces an imminent shortage of faculty for the arts and sciences, this issue has obviously declined in significance. See William Bowen and Julie Ann Sosa, 1989.

²The amendments to the ADEA also required a study of the implications of uncapping to be done by the National Academy of Sciences. However, that study was funded too late in the grace period to be able to conduct much quantitative research. The staff of the PFR and the National Academy study have communicated regularly to coordinate activities and reduce overlap.

³The full findings of the Project on Faculty Retirement are reported in Rees and my forthcoming study.

*Dean, College of Business Administration, Fordham University, Bronx, NY 10458. I draw freely from my joint work with Albert Rees. This research was supported by grants from the Andrew W. Mellon Foundation, the Carnegie Corporation of New York, and the William and Flora Hewlett Foundation. It also benefited from the expert research assistance of Jerene Good. I thank Rees for helpful suggestions on an earlier draft of this paper.

tions, articles and/or books published, research grants received, etc.) These analyses were based on two different sets of data. The first consisted of data on the current age of the tenured faculty and the flows into and out of tenure at a select set of institutions. The second consisted of a number of case studies and surveys conducted at a smaller number of institutions and surveys made by other organizations.

I. Format of the Study

I focus on the first concern of administrators: whether or not faculty will decide to retire when they are no longer required to do so. To analyze the impact of uncapping, however, it is necessary first to have a sense of faculty's current pattern of retirement. The PFR collected an original data set to do so. This data set was not, nor was it intended to be, a random sample. It consisted of data collected from a group of research and doctorate-granting universities and selective liberal arts colleges. It was designed to focus on the types of institutions in which either a change in retirement behavior was most likely to result from uncapping or a change in retirement behavior could lead to the most severe problems. The institutions where faculty were expected to delay retirement are prime research institutions. These are institutions in which faculty are most dedicated to research, the students are most qualified, teaching loads are light, and where active institutional association at least facilitates academic research and may be necessary for the research to occur at all. The institutions where delayed retirement could cause the greatest disruption are liberal arts colleges. These are institutions in which faculty sizes are so small that the presence of one faculty member who is incompetent could destroy the department.

Information was then gathered from thirty-three institutions (nineteen universities and fourteen liberal arts colleges) on the current age profile of tenured faculty (academic year 1988-89) by major discipline category (humanities, social sciences, and physical and biological sciences) and on the historical pattern of flows into (hire or pro-

motion)⁴ and out of (resignation, retirement, or death) tenure by discipline, type of flow, and age at the time of the flow. (Twenty-five institutions provided data for at least 5 years on all types of flows and the remainder provided data on retirements only.)

A number of states have already eliminated mandatory retirement by state law either for the tenured faculty in their state universities (for example, Connecticut, New York, Virginia) or for tenured faculty in all institutions of higher education (for example, Florida, Hawaii, Maine, Texas, Utah, Wisconsin) in their state. Wherever possible, data were collected for similar "capped" and "uncapped" institutions so that this "natural experiment" could be used to provide insight into the impact of ending mandatory retirement for all tenured faculty.⁵ Although some institutions uncapped very recently, others were uncapped for more than 10 years and can provide sufficient history to demonstrate whether or not the retirement decision changes when it is at the discretion of the faculty member. Data were gathered for tenured faculty in the arts and sciences only. (This limitation assured a more homogenous sample and was consistent with the primary interests of the PFR's principal source of funding.)

II. A Picture of Tenured Faculty

The data gathered on the current age distributions among tenured faculty in the sample provide a clue to the timing of when a change in retirement patterns (if it occurs as a result of uncapping) might be felt. The picture that emerges displays remarkable consistency regardless of type of institution, discipline, or capping status. The mean age

⁴In this context, "hire" signifies a hire with tenure to the institution in question and a "promotion" signifies an appointment to tenure at the institution from a previously untenured position. It should be noted that each of these flows represents a movement of individual faculty members into and out of individual institutions of higher education.

⁵There were insufficient selective private universities that were already uncapped to make this distinction for that category.

varies from 49.6 to 51.6. Although faculty are dispersed across an array of age cohorts, in most instances, they show the greatest concentration in the 46–50 age cohort. Thus the largest proportion of the tenured faculty is nearing but has not yet reached the age at which the retirement decision begins to come under consideration. Absent any move toward a permanent exemption from the ADEA for tenured faculty, they will make that choice voluntarily.

At the same time, the proportion of tenured faculty over age 70 is very small—0.7 percent in uncapped public universities is the largest.⁶ Moreover, less than 0.1 percent of all tenured faculty in the data set are over age 71. Indeed, in a data set of 8232 tenured faculty in all institutions, there are only 2 individuals who remain as active faculty members after age 75. The existence of a mandatory retirement law does not prevent the presence of active faculty beyond the mandatory age. The distinction is that in a capped environment, the decision to remain beyond the mandatory retirement age rests with the institution, whereas in an uncapped world, the decision to remain rests with the individual.

The proportion of faculty who choose to remain beyond age 70 depends in large part on the decisions that have been made at earlier ages. The 1978 amendments to the ADEA raised the mandatory retirement age for tenured faculty from 65 to 70 effective July 1, 1982. Thus sufficient time had elapsed for the faculty in this data set to have aged enough to take advantage of that change in the mandatory retirement age. Despite this opportunity, the proportion in the 66–70 age cohort ranged from 2.5 percent (capped liberal arts colleges) to 6.6 percent (private capped universities). Thus the proportion who will have the opportunity to choose to remain active beyond age 70 is likely to be quite small.

⁶The label of capped or uncapped pertains to the institution's status at the end of the period. Most of the institutions uncapped some time during the period of the data: one in 1976, one in 1978, three in 1980, three in 1984, one in 1986, and two in 1988.

Turning now to the data on flows, patterns emerge showing characteristic differences by type of institution. In liberal arts colleges, it appears that most faculty arrive early in their careers (without tenure) and if they subsequently receive tenure, they remain until retirement. In private universities, in contrast, faculty are nearly as likely to arrive with tenure as to be promoted to it. The pattern for public universities falls between these two.

An examination of the mean ages at which these events occur shows some surprising patterns. (However, the dispersion is such that none of the differences observed is statistically significant.) Hires with tenure, on average, take place at a later age (the early 40s) than promotions to tenure (the mid- to late 30s). Moreover, many hires take place at much later ages: in the private universities, 20 percent of these hires were age 50 or older. In fact, there are hires over age 60! These observations certainly call into question the assumption that retiring faculty would be replaced by newly minted, untenured, and relatively low-paid assistant professors. Indeed, an early retiree might be replaced by a newly hired but older and possibly more distinguished faculty member.

The data on retirement also show differences in mean age by type of institution. Mean retirement age at 66.8 is highest in private universities and lowest in liberal arts colleges at 65.3 for capped colleges and 64.3 for uncapped colleges—with public universities in between. There is generally little difference by discipline. In both public universities and liberal arts colleges, the mean age at retirement is generally slightly lower in the uncapped type.

The proportion of retirements at 70 or older also vary by institutional type and by individual institution within type. The proportion tends to be smallest in the liberal arts colleges (especially the uncapped ones, most of which show no late retirements) and largest in the private universities—where it is as much as 73 percent for one institution. Even within this institutional type, one institution had no late retirements and two had less than 5 percent.

TABLE 1—PATTERNS OF RETIREMENT
OF TENURED FACULTY

	(1)	(2)
Constant	-1.02 (-3.42)	56.23 (18.91)
SAT Scores per 100 Points of Combined Score	0.098 (4.14)	0.76 (3.24)
Research University	0.21 (4.11)	1.45 (2.84)
Formal Early and/or Phased Retirement Plan	-0.054 (-0.96)	-0.91 (-1.62)
Uncapped 5 years or More	-0.009 (-0.13)	0.014 (0.02)
Adjusted R-squared	0.56	0.40
No. Observations	31	31

Note: The *t*-values are shown in parentheses beneath regression coefficients.

Dependent variables: Col. (1) Proportion of retirements at age 70 or later; Col. (2) Mean age at retirement.

Regression analysis sheds further light on the differences in retirement ages. Results for two of the key regressions are displayed in Table 1. Each observation is an institution which provided 5 years of data on retirements. There were 31 observations used because not all data for all variables were available for all institutions. The dependent variable in the first regression is the proportion of retirements at age 70 or later (col. 1). Thus this regression does not address uncapping. Instead it provides insight into the determinants of "late" retirement. The dependent variable in the second regression is the mean age at retirement (col. 2). Thus this regression addresses the factors that tend to either postpone or expedite retirement.

The independent variables are the same for each regression. The first is the sum of the mean verbal and quantitative SAT scores of entering freshmen as reported in *Barron's Profiles of American Colleges* (1988). This variable is a direct measure of the quality of the students and an indirect measure of the quality of the institution. The variable is highly significant which suggests that professors are more likely to postpone retirement when they are teaching good students or are

in a good institution. The second independent variable takes the value one if the institution is classified as a research university,⁷ and zero otherwise. The coefficient of this variable is highly significant and of substantial size—suggesting that when a large part of a faculty member's job is research, retirement is likely to be postponed. Indeed these results suggest that the mean age at retirement is 1.45 years higher in research universities than elsewhere, all other things being equal. The third independent variable takes the value one if the institution has a *formal* early or phased retirement plan and zero otherwise. Although the coefficient of this variable has the right sign and is of substantial size, it is not significant at conventional levels—probably because of the small sample size. The fourth independent variable takes the value of one if the institution had no mandatory retirement age during the 5-year period covered by the data and zero otherwise. As expected, it has no impact in the first regression. The startling finding, however, is that this variable is zero and insignificant in the second regression as well.

These results suggest that tenured faculty retire later when their positions consist largely of research, their teaching loads are relatively light, and when they do teach, their students are good. Moreover, uncapping has had no perceptible effect on the mean age of retirement in selective liberal arts colleges and prominent public universities. Whether behavior might be different for private elite research universities or change over time in liberal arts colleges and public universities remains an open question.

In order to explore the implications of any change in retirement behavior in a totally uncapped world, the future composition of tenured faculties in each of the institutional types was projected using the

⁷This Carnegie classification includes those institutions that offer a full range of baccalaureate programs, give high priority to research, receive substantial federal support for research and development, and award at least 50 Ph.D. degrees a year.

Faculty Cohort Model⁸ and some very conservative assumptions concerning overall faculty growth and faculty retention rates—zero growth in the tenured faculty and a retention rate in the oldest cohorts that is 50 percent higher than that observed in currently uncapped institutions. The projections that result show little grounds for concern over a future dearth of opportunities for new faculty or for minority faculty. Instead the projections suggest an increase in the proportion of faculty age 40 or less by 2004, as other aspects of faculty demographics dominate any possible change in retirement behavior.

It appears, then, that there is little ground to suspect that large numbers of faculty will choose not to retire in a permanently and totally uncapped world. However, there is an exception to this conclusion: a few institutions that are presently still capped and whose faculty now tend to remain active until the mandatory retirement age of 70 will have substantial problems. In the PFR sample, these are elite private research universities.

There remains the concern, however, that the abilities of those who do remain will be much reduced with age. Here again there is little evidence to support this fear. Although research activity declines from mid-career, it does not cease. (See Alan Howe and myself, 1990.) The outlook for teaching effectiveness varies by discipline, but does

not suggest that age has a large effect in any case. (See Daniel Kinney and myself, 1989.) The evidence thus suggests that administrators' fears concerning the ending of mandatory retirement in the arts and sciences are largely groundless for much of higher education. The few faculty members who do choose to remain active beyond age 70 are unlikely to provoke questions of competence. Moreover, personnel policies can be designed that will encourage these faculty members to select retirement.

REFERENCES

- Biedenweg, Rick and Thomas Keenan, *The Faculty COHORT Model User Manual*, Stanford University, 1989.
- Bowen, William G. and Julie Ann Sosa, *Prospects for Faculty in the Arts and Sciences*, Princeton: Princeton University Press, 1989.
- Howe, Alan B. and Sharon P. Smith, "Age and Research Activity," unpublished paper, Industrial Relations Section, Princeton University, 1990.
- Kinney, Daniel P. and Sharon P. Smith, "Age and Teaching Performance," unpublished paper, Industrial Relations Section, Princeton University, 1990.
- Rees, Albert and Sharon P. Smith, *Faculty Retirement in the Arts and Sciences*, Princeton: Princeton University Press, forthcoming 1991.
- Rosovsky, Henry, *The University: An Owner's Manual*, New York: W. W. Norton, 1990.
- Barron's *Profiles of American Colleges*, 16th ed., New York: Barron's Educational Series, 1988.
- Carnegie Foundation for the Advancement of Teaching, *A Classification of Institutions of Higher Education*, Princeton: Carnegie Foundation Technical Report, 1987.

⁸This faculty planning model is a Markov Chain Model with feedback. The model was written by Rick Biedenweg and Tom Keenan for use at Stanford University. The Consortium on Financing Higher Education is making it available for use at any institution. See Biedenweg and Keenan (1989) for an explanation of the workings of this model.

The Effects of Pensions and Retirement Policies on Retirement in Higher Education

By ALAN L. GUSTMAN AND THOMAS L. STEINMEIER*

Exits through retirement are one of the important labor market flows that shape the age structure of faculty, and help to determine the quality and costs of higher education. As a result of perceived pressures from demographic changes, as well as in reaction to a series of policy initiatives, including a legal requirement that mandatory retirement be eliminated in higher education after 1993, institutions of higher education have become increasingly interested in predicting and influencing retirement behavior of faculty, while exerting more control over the associated costs of compensation.

Life cycle, structural econometric models of retirement have been used to analyze analogous policy issues at the national level (Gary Fields and Olivia Mitchell, 1984; ourselves, 1986a; Robin Lumsdaine et al., 1990). These models first specify and estimate the various components of the opportunity set. Then, based on the subsequent retirement behavior of those facing alternative opportunities, the parameters of the utility function that underlie the retirement decision are estimated. Once the components of a structural retirement model are estimated, effects on retirement outcomes of policy changes may be simulated (see our 1985, 1986b, and forthcoming articles).

In the present paper, we apply these techniques to analyze retirement behavior of the tenured, male faculty employed or retiring in the late 1970's at 26 member

colleges and universities of the Consortium On Financing Higher Education (COFHE), a group that includes some of the highest quality private colleges and universities in the country. The estimated model is used to analyze the effects of early retirement incentives and changes in mandatory retirement rules. Although the sample is not fully representative of higher education, and the data are over a decade old, these are the best data available for illustrating the usefulness of the recent developments in retirement research in analyzing related behavior and policies in higher education.

I. The Empirical Specification and Data

The utility function to be estimated is CES and is given by

$$U = (1/\delta) \int_0^T \{ [C(t)]^\delta + e^{X_t\beta + \varepsilon} [L(t)]^\delta \} dt$$

where $C(t)$ and $L(t)$ are consumption and leisure at time t , and T is the relevant time horizon; X_t includes age and a constant that affect the relative weight of leisure in the utility function at time t , and β is the associated vector of parameters, presumed to be constant across both time and individuals; δ (with $\delta \leq 1$) and ε are time-invariant stochastic terms reflecting, respectively, the elasticity of substitution between consumption and leisure for each individual and the relative weight that the individual places on leisure.

This utility function is maximized with respect to consumption and leisure, subject to the lifetime budget constraint

$$\int_0^T e^{-rt} \{ y[L(t), t] - C(t) \} dt = 0$$

where $y[\cdot]$ is the function relating compensation to leisure (and hence to work effort), and r is the real interest rate. The compen-

*Dartmouth College, Hanover NH 03755 and NBER, Cambridge, MA 02138; and Texas Tech University, Lubbock, TX 79409, respectively. We are grateful for data provided by Katherine H. Hanson of the Consortium On Financing Higher Education, to Euysung Kim, Scott Miller and Fony Suryapranata for research assistance, the Department of Education for research support, and Steven Venti, Olivia Mitchell and W. Lee Hansen for comments. This paper is part of the NBER programs in Labor Studies and Aging. Any opinions expressed are solely our own.

sation function reflects the effects of wages, pensions, and Social Security, with the effects of pensions and Social Security calculated as the difference in the present value of benefits attributable to additional work.

The data used in the structural retirement analysis pertain to arts and sciences faculty employed by these schools on December 31, 1978, to retirees from these schools from the 1973–74 academic year through the end of 1978, and to the pensions and retirement programs reported by these schools in August 1979. For each faculty member employed over the covered period, the following information is provided in the survey: date of birth; sex; date of entry into tenured track status; indicators of tenure status; date of tenure; 9-month salary rate; full or part-time status; date of retirement, departure or termination, if appropriate; whether, in the case of a departure, it was due to mandatory retirement; and other information about the faculty member.

Importantly, however, no information is available on health status. Each school is identified. For the 1979 period, pension plan provisions, early retirement and flexible retirement policies are reported for each school in separate COFHE documents. (These data have also been used by John Blackburn and Susan Schiffman, 1980; and J. Russell Southworth and Ronald Jagmin, 1979.) For later simulations, we have obtained descriptions of these plans for 1989 directly from the schools.

In constructing the opportunity set, wage equations for full and part-time work are estimated. (These equations are in a statistical appendix available upon request.) Wages are projected using the experience and tenure coefficients from the wage equations, assuming a general wage growth equal to the growth of average hourly earnings nationally. Social Security benefits are calculated on the basis of the rules that were applicable to each cohort (see our 1985 article). The calculations include retired worker benefits, spouse benefits, and survivor benefits. Required faculty pension contributions are subtracted from wages.

In 1979, normal retirement (NR) age was 65 in 18 of the COFHE schools in the

sample, 66 in 1, 68 in 5, and 70 in 2. From 1979 to 1989, the median normal retirement age remained at 65, but 5 schools reduced the normal retirement age from 70 or 68 to 65. Mandatory retirement (MR) age was 65 in 5 schools, 68 in 5 schools, 70 in 9 schools, and 7 did not report a mandatory retirement age. By 1989, mandatory retirement age was 70 at all 26 schools. Separate information is also provided on the actual application of mandatory retirement to each of the cases in the sample (Southworth-Jagmin). The fractions of retirements in the sample which were mandatory are 78, 99, and 97 percent for those 65 to 67, 68 to 69, and 70 to 72 years-old, respectively.

At 7 schools, early retirement (ER) supplements were generally available in 1979, and at 4 schools they were available on an *ad hoc* basis. The availability of these programs has spread rapidly over time. For the 1979 sample of 26 schools, at the plan's early retirement age, pension wealth amounts to 5.9 times earnings. At normal retirement age, the wealth-yearly earnings ratio rises to 6.0 to 1, and at mandatory retirement age, pension wealth averages 6.8 times yearly earnings. By 1989, formal early retirement programs were available at 23 of 29 schools reporting to COFHE, with *ad hoc* programs at the other 6.

Table 1 highlights the accrual of pensions and any early retirement bonuses. In calculating the accrual profiles, the table uses the average birth date and wage profiles by institution and then weights the results by the number of tenured faculty at each institution. It can be seen from Table 1 that even though we are dealing with defined contribution plans, early retirement programs available at the schools create significant spikes in the accrual profile.

The preretirement period (PR) is the 5 years preceding the year before eligibility for early retirement benefits. The ER spike is computed over the year in which eligibility for early retirement benefits is obtained, and the NR spike is computed over the year before eligibility for normal retirement benefits is obtained. Early Retirement (ER) is the period between the periods for computing the ER spike and NR spike. Finally, Late Retirement (LR) is the period be-

TABLE 1—INCREMENT IN PENSION WEALTH AND
EARLY RETIREMENT BONUS/EARNINGS

Age at Hire	PR	ER Spike	ER	NR Spike	LR
Early Retirement Available					
25	15.4%	149.6%	-19.5%	-10.1%	0.1%
35	15.4	154.3	-19.9	-10.8	0.1
45	15.2	80.2	0.4	1.6	0.5
No Early Retirement Available					
25	14.1%			21.1%	8.2%
35	14.1			21.1	8.2
45	14.1			34.5	8.2

Note: PR: preretirement period; ER: early retirement; NR: normal retirement; LR: late retirement.

tween attaining eligibility for normal retirement benefits and mandatory retirement.

The sharp increments in the accrual rates reflect the attainment of eligibility for early or normal retirement benefits, where eligibility is accompanied by a bonus. In one plan, for example, those retiring early receive 60 percent of yearly salary between ages 62 and 65. On average, covered individuals hired at age 35 become eligible for a supplement worth 1.5 years of salary when they reach the early retirement age.

The survey also provides information on whether the individual was working part time at the end of the survey or at the date of retirement, if earlier. If more than 10 percent of retirements from a school are from part-time work, partial retirement is assumed to be generally available at the school.¹

Consider now the structure of the dependent retirement variable. For many of those in the sample, it is possible to determine a sequence of outcomes for 6 years. For those who have retired, there is information on status just before retirement. However, the sample of retirees from COFHE schools is not a true panel. For those who were partially retired as of December 31, 1978, the duration of partial retirement is not reported. Information is not provided indicating who partially retired after having left a state in which they worked full time and

TABLE 2—PARAMETER ESTIMATES
OF UTILITY FUNCTION

γ	Parameter: $F(\delta) = e^{\gamma(\delta-1)}$	0.19 (77.39)
σ_ϵ	Standard Deviation of ϵ	2.17 (18.27)
ρ	Parameter: $E(\epsilon \delta) = \rho(1-\delta)$	-4.46 (19.78)
β_0	Constant in β	7.81 (12.40)
β_1	Coefficient of (Age-62) in β	0.26 (5.01)
Observations of Vintage 1909-12 Individuals		273
Number of Weighted Observations		337
Log Likelihood		-407.78

whose lifetime job at the university involved part-time work. In an effort to distinguish partial retirees who reduced work effort from full time from those who were only part-time employees during prime working age, the sample includes only tenured faculty. An examination of the frequencies of retirement and partial retirement between the ages 40 and 60, indicates that early leaving and part-time work by tenured faculty are not important in the sample. (The frequency distribution of retirement sequences for the 6 years covered by the survey are reported in an Appendix and in a report to the Department of Education, both available on request.)

II. Empirical Findings

Parameter estimates from the model are presented in Table 2, with asymptotic standard errors indicated in parentheses below each figure. γ and ρ are parameters of the distribution of δ and ϵ . The data were fit to males in cohorts born from 1909 through 1912. The estimation procedure follows our article (1986a), with one modification. Those who retired before 1974 were not included in the sample. Accordingly, the estimation procedure attempts to correct for selection due to prior retirements using early retirees from younger cohorts, a correction that was not required in our earlier work. The parameters are significant at standard levels.²

¹Although our estimates assume that partial retirement involves a reduction in official hours at work, it is also possible that minimum hours constraints may be less binding in academia than elsewhere.

²There were 5 observations in which individuals retired the year before becoming eligible for early retirement bonuses worth an additional 1 or 2 year's

TABLE 3—ACTUAL AND SIMULATED
EMPLOYMENT PERCENTAGES

Age	Actual		Simulated	
	FT	PT	FT	PT
62-64	87.6%	3.9%	76.6%	5.9%
65-67	38.9	15.0	38.6	7.1
68-69	12.7	12.7	17.2	6.2
70-72	1.6	5.6	6.2	4.0

Note: FT: full-time work, PT: part-time work.

COFHE faculty retired later than did workers covered by the Retirement History Survey. By age 64, 39 percent of healthy males in the survey without a pension had left full-time work, while only 14 percent of all faculty in the COFHE sample had. These differences are reflected in differences in the estimated constant, but the coefficients estimated for the effects of aging are similar for the two samples. This suggests that although, for any given opportunity set, the levels of retirements are higher in the COFHE sample, the changes in retirement rates induced by a given incentive will be similar.

Simulation is accomplished by applying Monte Carlo techniques. For each observation, five random draws are taken for the stochastic terms δ and ε in the utility function. Table 3 simulates the percentages working full and part-time by age using the pensions and Social Security rules in place during the sample period. The percentages resulting from a simple nonparametric hazard model are presented for comparison. The model simulates the substantial drop in work effort during the period fairly well, although it has some trouble capturing the large increase in part-time work after age 65.

salary. In a model such as the one used in this paper, such an event would occur only if the coefficient on age in the utility function were implausibly high, implying almost no response of retirement behavior to economic incentives. Since these 5 retirements may well be due to health problems that, in the absence of information on health status, cannot be controlled for, and since these observations would dominate the results if they were included, these 5 observations are excluded from the sample.

TABLE 4—EFFECTS OF RAISING MANDATORY
RETIREMENT AGE

Age	Percentages Working Full Time with 1979 Pensions		
	Actual MR	MR = 70	No MR
65-67	38.6%	65.6%	64.7%
68-69	17.2	53.8	53.0
70-72	6.2	6.2	43.4

Note: MR: mandatory retirement.

Table 4 presents the results of simulations that raise the mandatory retirement age, holding other aspects of the compensation profile the same. The first column uses the actual mandatory retirement ages observed in the sample, while the following two columns raise the minimum mandatory retirement age to 70 and eliminate it, respectively. Reflecting the frequency with which individuals worked until mandatory retirement, these simulations suggest massive increases in full time work by faculty members in their late 60s and early 70s. As a word of caution, the simulations in column 3 take us outside the age range of observation for those in the sample. For that reason, and because we do not have information on health status, those results should be interpreted with some care.

Using the observed 1989 pension plans, all of which specified age 70 as the mandatory retirement age, we also simulated the effects of early retirement supplements. Compared to a situation where the early retirement supplements are suppressed, the observed supplements for each school reduce the percent working full time from 66.7 to 65.9 percent for 65 to 67 year-olds, and from 55.1 to 53.7 percent for 68 to 69 year-olds. Therefore, if the early retirement provisions in the pensions available in 1989 were abolished, there would be little overall effect on observed early retirement behavior.

We also simulated the effects of uniformly adopting one school's 1989 early retirement plan, a plan that provides up to 40 percent (1.33 percent per year of service) of salary either up to 5 years or until age 70, whichever is sooner. For this plan we find that full-time work is reduced to 64.4 per-

cent for 65 to 67 year-olds, and to 51.0 percent for 68 to 69 year-olds, a relatively small effect. Thus early retirement incentive plans do not appear to be very effective, at least in the COFHE schools.

When the costs of the retirement plan are simulated, three elements are of importance: the rent paid to those who would have retired at younger ages on the basis of unchanging retirement behavior; the reduction in costs due to accelerated retirement by some highly paid faculty; and the costs of replacement faculty. Our calculations (shown in the Appendix) indicate that rents exceed the savings due to early retirement, so that even ignoring the costs of replacement faculty, this type of early retirement incentive plan will not be cost saving.

III. Conclusions

The data used in this study are old, are not strictly longitudinal, are missing key pieces of information on health status, family structure, field of specialization, and postretirement behavior outside the primary employer, and apply only to a limited and not representative sample of colleges and universities. These limitations mean that the findings from this study should be applied with caution. Nevertheless, the preceding analysis indicates the feasibility of adapting recent innovations in the retirement literature for analysis of retirement policies by institutions of higher education.

Simulations suggest that for the COFHE schools, extending and then eliminating mandatory retirement will lead to a significant number of faculty to postpone retirement.

Some institutions of higher education are considering early retirement incentive programs that will have costs and benefits that are very sensitive to the induced retirement responses. For these plans to be cost saving, the savings from inducing earlier retirement by higher-paid senior faculty must exceed the costs from rents accumulating on the basis of unchanging retirement behavior and replacement costs. Our calculations suggest they will not. An obvious alternative option that might be considered in an effort to

influence faculty retirement is the adoption of a defined-benefit plan, that is offered at many public institutions of higher education, and that can create even stronger early retirement incentives.

The effects of retirement incentives created by innovative retirement programs, the associated program costs, and implications of related policy initiatives, may all be analyzed with analytical tools that are currently available. All that is required is the availability of the required data.

REFERENCES

- Blackburn, John O. and Schiffman, Susan**, "Faculty Retirement at The COFHE Institutions: An Analysis of The Impact of Age 70 Mandatory Retirement and Options For Institutional Response," Washington: Consortium On Financing Higher Education, May 1980.
- Fields, Gary S. and Mitchell, Olivia**, *Retirement, Pensions and Social Security*, Cambridge: MIT Press, 1984.
- Gustman, Alan L. and Steinmeier, Thomas L.**, "The 1983 Social Security Reforms and Labor Supply Adjustments of Older Individuals in the Long Run," *Journal of Labor Economics*, April 1985, 3, 237-53.
- _____, and _____, (1986a) "A Structural Retirement Model," *Econometrica*, May 1986, 54, 555-84.
- _____, and _____, (1986b) "A Disaggregated Structural Analysis of Retirement By Race, Difficulty of Work and Health," *Review of Economics and Statistics*, August 1986, 67, 179-85.
- _____, and _____, "Changing The Social Security Rules For Workers Over 65," *Industrial and Labor Relations Review*, forthcoming.
- Lumsdaine, Robin L., Stock, James H. and Wise, David A.**, "Three Models Of Retirement: Computational Complexity vs. Predictive Validity," xeroxed, 1990.
- Southworth, J. Russell and Jagmin, Ronald A.**, "Potential Financial and Employment Impact of Age 70 Mandatory Retirement Legislation On COFHE Institutions," Washington: Consortium On Financing Higher Education, 1979.

Conflict and Attitudes Toward Risk

By STERGIOS SKAPERDAS*

The outcome of conflict is rarely known with certainty before it occurs. This is true even when those involved have perfect information about each other's capabilities and idiosyncracies. In military contexts there is enough residual uncertainty in the battlefield so that the side with the superior force cannot be guaranteed to win. Similarly, in cases of economic warfare (trade wars, battles for market share, rent-seeking situations), greater expenditure of resources than one's opponent does not necessarily yield a higher *ex post* payoff. Consequently, one would expect attitudes toward risk to influence the strategic choices made by parties that are about to engage in conflict and, through these choices, the likelihood of winning.

This paper explores the effect of differential attitudes toward risk in a model of conflict.¹ Is greater risk aversion than one's opponent advantageous or not, and does increasing risk aversion reduce or increase the resources expended on conflict? These are the two main questions examined under two alternative institutional arrangements.

I. When Conflict is Inevitable

The model employed here is a variation on models analyzed by Jack Hirshleifer

[†]*Discussants:* Michelle R. Garfinkel, Federal Reserve Bank of St. Louis; Martin C. McGuire, University of Maryland.

*Department of Economics, University of California, Irvine, CA 92717. I am grateful to Chew Soo Hong, Ami Glazer, Jack Hirshleifer, Dan Klein, and Klaus Nehring for helpful comments.

¹The role of risk in conflict is a recurring theme in Thomas Schelling's work (1966, see ch. 3). In other related work, John Gould (1973) has explored the role of risk aversion in producing out-of-court settlements in legal conflicts.

(1988) and by myself (1990). There are two parties, labeled 1 and 2, each possessing one unit of an initial resource. The decision facing each party is how much of the resource to convert into arms, leaving the remainder for useful production. Letting y_i denote the quantity of arms chosen by party i ($= 1, 2$), $1 - y_i$ represents the party's remaining productive resource. The relative quantity of arms determines the probability of each party to win in conflict. The winner receives a prize that depends on the remaining productive resources of both parties. The prize is a function $C(1 - y_1, 1 - y_2)$ which is increasing in both arguments; the loser receives nothing. Let $p(y_1, y_2)$ be a function (increasing in y_1 and decreasing in y_2) that represents the win probability of party 1, while $1 - p(y_1, y_2)$ represents the win probability of party 2. In addition, for a given (y_1, y_2) pair we have $1 - p(y_1, y_2) = p(y_2, y_1)$ so that having equal win probabilities is equivalent to $y_1 = y_2$, and a party has higher win probability than her opponent if and only if she has a greater quantity of arms. Both parties are expected utility maximizers and 2 is *strictly more risk averse* than 1 in the Arrow-Pratt sense. Then, with $U(\cdot)$ and $V(\cdot)$ being the $VN - M$ utility functions of 1 and 2, respectively, there exists an increasing and strictly concave function $k(\cdot)$ such that $V(\cdot) = k(U(\cdot))$. Letting $V(0) = U(0) = 0$, the expected payoffs for a given choice of arms allocations, (y_1, y_2) , are

$$(1a) \quad \pi^1 = pU(C)$$

$$(1b) \quad \pi^2 = (1 - p)V(C) = (1 - p)k(U(C))$$

where $p \equiv p(y_1, y_2)$ and $C \equiv C(1 - y_1, 1 - y_2)$. While the prize the winner receives depends on the strategic choices of both

parties and is thus endogenous, the loser's payoff is independent of the strategies adopted. The latter feature is partly responsible for the unambiguous results obtained below.

The first objective is to find out whether the less risk-averse side has a higher or lower win probability in equilibrium than her opponent. In order to isolate the effect of differential attitudes toward risk, I consider a *symmetric* function $C(\cdot, \cdot)$ in the sense that for any z and w , we have $C(z, w) = C(w, z)$, implying that $C_1(z, w) = C_2(w, z)$ where the subscript 1 (2) denotes the partial derivative with respect to the first (second) argument of $C(\cdot, \cdot)$. (Otherwise, $C(\cdot, \cdot)$ has the properties of non-increasing marginal products and a non-negative cross derivative; within these constraints, returns to scale can be increasing, constant, or decreasing.) The two parties choose their strategies, y_1 for 1 and y_2 for 2, simultaneously. Attention is restricted to interior Nash equilibria where both sides choose a positive quantity of arms. Then, by differentiating (1a) with respect to y_1 and (1b) with respect to y_2 , and by setting the derivative equal to zero, at any equilibrium point, we have

$$(2a) \quad p_1 U(C) = p U'(C) C_1$$

$$(2b) \quad -p_2 k(U(C)) \\ = (1-p) k'(U(C)) U'(C) C_2$$

where the subscripts 1 and 2 denote partial derivatives with respect to the first and second arguments of the relevant function (no special notation is used here for the equilibrium values of the relevant functions and variables). The left-hand side of each equation represents the marginal benefit (the change in the win probability times the utility of the prize) associated with a small increase in a party's quantity of arms, while the right-hand side represents its marginal cost. Dividing (2a) by (2b) and rearranging yields

$$(3) \quad (C_1/C_2) k(U(C)) \\ = k'(U(C)) U(C) \frac{p_1(1-p)}{-p_2 p}.$$

The strict concavity of $k(\cdot)$ along with the normalization $k(U(0)) = U(0) = 0$ implies that for any $C > 0$ we have

$$(4) \quad k(U(C)) > k'(U(C)) U(C)$$

and its application in (3) yields

$$(5) \quad C_1/C_2 < p_1(1-p)/(-p_2 p).$$

Suppose, for a moment, that the less risk-averse party (1) has a higher win probability in equilibrium so that $p > 1/2$ and $y_1 > y_2$. Since $1 - y_1 < 1 - y_2$, it follows that $C_1 \equiv C_1(1 - y_1, 1 - y_2) \geq C_1(1 - y_2, 1 - y_1)$. The symmetry of the production function also implies that $C_1(1 - y_2, 1 - y_1) = C_2(1 - y_1, 1 - y_2) \equiv C_2$ which, by the last inequality, yields $C_1 \geq C_2$. Therefore, the right-hand side of (5) is strictly greater than 1. Under a rather weak assumption, however, which also guarantees existence of equilibrium,² it can be shown that the right-hand side of (4) is greater than one if and only if $p < 1/2$. Hence, the initial supposition of $p > 1/2$ is contradicted and we must have $p < 1/2$ at any interior equilibrium. In conclusion, the more risk-averse party always has a higher win probability in equilibrium and thus being *more risk averse than one's opponent is advantageous when conflict is inevitable*.³ The quantity of arms invested by each party can be thought of as insurance against losing in conflict. Since the more risk-averse party can also be thought of as more fearful of losing, that party will also invest more in arms and thereby have a higher win proba-

²The assumption is that $p_{11}p < p_1^2$ and, equivalently, $-p_{22}(1-p) < p_2^2$ where the subscripts denote derivatives with respect to the first and second arguments of $p \equiv p(y_1, y_2)$. This assumption is satisfied by any p function that is concave in its first argument as well as for others that are not "too convex." The proof of this result and of existence of equilibrium can be found in my earlier paper. It should be noted that existence of equilibrium does not necessarily preclude the two sides from being risk lovers; the production function could compensate for the nonconcavity of the utility functions.

³This result does not depend on having a unique pure-strategy equilibrium. If multiple interior equilibria were to exist, the result would hold for every single one of them.

bility.⁴ The intuition for the second result of this section, to which I now turn, is similar.

To consider the effect of increasing risk aversion on the total amount of resources expended on conflict, two simplifying assumptions are introduced. First, the two parties are now identical in their risk attitudes in addition to the model's other attributes. Thus, we seek to compare a situation where the utility function of both parties is described by $U(\cdot)$ to that where the utility function is $k(U(\cdot))$. Second, there is a unique equilibrium that together with the assumption of identical parties implies that the equilibrium must be symmetric and, therefore, the two sides choose the same quantity of arms and they have equal win probabilities. When the utility function is $U(\cdot)$ (respectively, $k(U(\cdot))$), the reaction function of either party is represented by $r^u(\cdot)$ (respectively, $r^v(\cdot)$) and the equilibrium quantity of arms is denoted by y^u (respectively, y^v). For each $i = u, v$ we have $r^i(y^i) = y^i$ and $r^i(0) > 0$ (otherwise, $(0, 0)$ would be an equilibrium). Then, since y^i is a unique and symmetric equilibrium, it must be the case that $r^i(y) > y$ for all $y \in [0, y^i]$. It can also be shown, using the first-order conditions along with (4), that $r^v(y) > r^u(y)$ for all $y \in (0, 1)$. If we had $y^v \geq y^u$, the two just-derived inequalities would imply that $y^v \leq r^u(y^v) < r^v(y^v)$, which contradicts $r^v(y^v) = y^v$. Hence, we must have $y^v > y^u$ and, therefore, *as the two sides become more risk averse, the amount of resources expended on conflict increases.*

II. When Settlement is Possible

Up to this point it has been assumed that open conflict is inevitable, but this need not

be the necessary outcome of interaction between the two parties even after both have armed themselves. Any simultaneous choice of arms determines each side's win probability and the size of the prize, C , that the winner receives. Now suppose that the two parties can communicate and are able to divide the prize in any way they agree. An agreement can be defined as a number β (between 0 and 1) where β represents the share of the prize received by party 1 with the remainder, $1 - \beta$, going to party 2. Then, the payoffs under some agreement β would be as follows (where the utility functions are as in (1)):

$$(6a) \quad \pi^1(\beta) = U(\beta C)$$

$$(6b) \quad \pi^2(\beta) = V((1 - \beta)C) \\ = k[U((1 - \beta)C)].$$

For an agreement to be credible, it would be necessary for both parties to be at least as well off as under open conflict. Otherwise, the party with lower payoff under the agreement could declare war. Formally, an agreement β is *credible* if and only if $\pi^i(\beta) \geq \pi^i$ for both $i = 1, 2$ where π^i is as defined in (1). If both parties were risk averse, it can be shown easily that the range of credible agreements would be an interval $[\beta, \bar{\beta}]$ such that $U(\beta C) = pU(C)$ and $k[U((1 - \bar{\beta})C)] = (1 - p)k(U(C))$. Note that this range of agreements depends on p and C , with both determined by the simultaneous strategic choices of the two parties, and on the utility functions. Since with risk aversion we have $U(pC) > pU(C)$ and $k[U((1 - p)C)] > (1 - p)k(U(C))$, the division of the prize according to the win probabilities belongs to the set of credible agreements.

With both parties being risk lovers, however, the set of credible agreements is empty; the two inequalities stated above are reversed and β is always strictly less than $\bar{\beta}$. In this case conflict would always take place. When one party is risk averse and the other one is a risk lover, credible agreements may not exist depending on the specific utility functions and the other parameters of the model.

⁴Alternatively, it could be argued that the more risk-averse party would be less willing to "gamble" on winning the conflict and thus be at a disadvantage. This intuition, which is completely opposed to the one materialized in my model, is not totally misplaced and it leads to ambiguous effects in the single-agent analog to the problem explored here, the problem of self-protection (see Georges Dionne and Louis Eeckhoudt, 1985, and Martin McGuire et al., 1991). As mentioned earlier, the unambiguous result obtained here is partly due to the fact that the loser's payoff is independent of the amount of arms investment.

For the remainder of this section it is assumed that both parties are risk averse, leaving a nonempty set of credible agreements. For any given choice of arms (y_1, y_2) , the two parties face a bargaining problem with the payoffs in (1) representing the disagreement payoffs and a payoff (or, utility) possibilities set whose frontier can be derived from (6). In terms of shares of the prize C , an agreement could be reached anywhere in the interval of credible agreements $[\beta, \beta]$. Without a determinant bargaining outcome, a rule assigning to each (y_1, y_2) pair a unique β , no results about the effect of risk attitudes are forthcoming. But suppose for the moment that the outcome of bargaining is such that the prize is always divided according to the win probabilities with parties 1 and 2 receiving pC and $(1-p)C$, respectively. Then, the payoff functions in (6) would be $\pi^1(p)$ and $\pi^2(p)$. Note that p and C are functions of (y_1, y_2) , and although conflict will not occur, each side invests in arms so as to balance the marginal benefit of a better bargaining position against the marginal cost associated with the reduction in the value of the prize. At an interior equilibrium point, the following first-order condition must be satisfied for 1:

$$(7) \quad U'(pC)[p_1C - pC_1] = 0$$

implying that $p_1C - pC_1 = 0$, which does not depend on the utility function. Instead, party 1 (and by a similar argument, party 2) behaves as if he were risk neutral. In addition, it can be shown that in this case the two sides of (5) are equal and that the prize would be split in half. Thus, *in this special case about the outcome of negotiations, attitudes toward risk do not have any influence on the strategic choices of the two parties.* Furthermore, since identical risk-averse parties would expend greater resources under conflict than identical risk-neutral parties would, and the two parties under the settlement arrangement behave as if they were risk neutral, fewer resources are wasted when settlement is possible than when it is not.

Although setting the bargaining agreement equal to p is the simplest possible rule

the two parties could follow as it is a natural "focal point," it has drawbacks. If party 1 were risk neutral and party 2 were risk averse, we would have $p = \beta$ which is the worst possible agreement for party 1. This example suggests a bias of this rule in favor of the more risk-averse party and is indicative of the possible inconsistencies one might discover in other examples. An alternative approach would be to employ an axiomatic bargaining solution, like the Nash or the Kalai-Smorodinsky solutions (Alvin Roth, 1979, contains a comprehensive analysis of this topic) that satisfy a set of properties (axioms) and are thereby consistent across different situations. Both of these solutions have been shown to favor the least risk-averse side (see Roth) when the disagreement point and the bargaining possibilities are deterministic, but the effect of employing either solution in the model of this paper is an open question.

III. Conflict or Settlement?

If at least one party is a risk lover, the set of credible agreements may be empty, implying conflict. The absence of the necessary communication channels for facilitating settlement is a second reason for entering into conflict. In turn, lack of communication could be due either to exogenous factors outside the control of the two parties, or to a prior strategic move by just one party. Obviously, after armament levels have been chosen, and in the presence of credible agreements, the incentive not to communicate or purposely destroy the channels of communication would not exist. Only in the case where one party would be better off if there were no possibility of communication (i.e., the payoff for that party would be higher under the institutional arrangement in Section I than that in Section II) would there be an incentive to make in advance an irrevocable commitment not to communicate when the time, and the temptation, to do so comes. This commitment must take the form of a "burn-the-bridges" act for the opponent to be convinced of its irrevocability.

What are the conditions that would make such a commitment profitable, or, in other

words, when would one side prefer the institutional arrangement of Section I? An answer to this question would depend on the particular assumptions employed about the outcome of bargaining in Section II. Some insight can be gained by considering the special case where, if the parties were to settle, they would agree on dividing the prize in accordance with the win probabilities. Recall that in this case, both sides would behave as if they were risk neutral which implies, given the symmetry of the model, that the prize would be split in half. In addition, recall from Section I that the less risk-averse party has lower win probability (less than $1/2$) when conflict is inevitable, and since the prize would be smaller as well under this arrangement, it follows that the less risk-averse party would always prefer the arrangements where settlement is possible and thus never commit to conflict. Although the prize would be smaller under conflict, the fact that the more risk-averse party has greater win probability under it implies that the settlement arrangement would not be necessarily preferable and, thus, if commitment to conflict were possible, it could only be optimal for the more risk-averse party.

Note that if the equilibrium under the settlement arrangement were to favor the least risk-averse party, the possibilities for the more risk-averse party instigating conflict by committing to it in advance would increase. Overall, a settlement and bargaining arrangement known to favor one party

runs the risk of driving the other one to conflict. The avoidance of strategic uncertainty about the outcome of negotiations, an important but difficult-to-model issue left out thus far, is another possible reason for either party, especially the more risk-averse one, to commit to conflict.

REFERENCES

- Dionne, Georges and Eeckhoudt, Louis, "Self-Insurance, Self-Protection and Increased Risk Aversion," *Economics Letters*, Nos. 1-2, 1985, 17, 39-42.
- Gould, John P., "The Economics of Legal Conflicts," *Journal of Legal Studies*, June 1973, 2, 279-300.
- Hirshleifer, Jack, "The Analytics of Continuing Conflict," *Synthese*, August 1988, 76, 201-33.
- McGuire, Martin, Pratt, John and Zeckhauser, Richard, "Paying to Improve Your Chances: Gambling or Insurance?," *Journal of Risk and Uncertainty*, forthcoming, 1991.
- Roth, Alvin E., *Axiomatic Models of Bargaining*, Lecture Notes in Economics and Mathematical Systems, Vol. 170, New York: Springer-Verlag, 1979.
- Schelling, Thomas C., *Arms and Influence*, New Haven: Yale University Press, 1966.
- Skaperdas, Stergios, "Cooperation, Conflict and Power in the Absence of Property Rights," mimeo., University of California-Irvine, May 1990.

The East European Revolution of 1989: Is It Surprising that We Were Surprised?

By TIMUR KURAN*

Many aspects of the East European Revolution are controversial, but on one point everyone agrees: it caught the world by surprise. Even local dissidents were stunned by the sudden turn of events.

We will never know how many East Europeans *did* foresee the explosion of 1989. But at each step, accounts painted a picture of nations united in amazement. To my knowledge, only one study addresses the issue systematically. Four months after the breaching of the Berlin Wall, the Allensbach Institute asked a broad sample of East Germans: "A year ago did you expect such a peaceful revolution?" Only 5 percent answered "yes," though 18 percent responded "yes, but not that fast." Fully 76 percent admitted to being totally surprised. These figures are all the more remarkable given the I-knew-it-would-happen fallacy—the human tendency to exaggerate foreknowledge (Baruch Fischhoff and Ruth Beyth, 1975).

Yet in hindsight the revolution appears as inevitable. In each of the six countries the leadership was despised, economic promises remained unfulfilled, and basic freedoms existed only on paper. More importantly, winds of change in the Soviet Union were making Soviet intervention increasingly unlikely. But if the revolution was indeed inevitable, why was it not foreseen? What kept us from noticing signs that now, after the fact, are so plainly visible?

I. Preference Falsification and Revolutionary Bandwagons

Consider a country featuring two camps competing for political power: government and opposition. Members of society, indexed by i , all place themselves publicly in one camp or the other, although a person may privately feel torn between the two camps. I am thus distinguishing between an individual's *private preference* and *public preference*. The former is effectively fixed at any instant, the latter a variable under his or her control. When his two preferences differ the individual is engaged in *preference falsification* (see my 1990a article).

Let S represent the size of the public opposition, expressed as a percentage of the population. Initially it is near 0, implying that the government commands almost unanimous public support. As a mass-supported seizure of political power, a revolution may be treated as an enormous jump in S .

Now take a citizen who wants the government overthrown. The likely impact of his own public preference on the government's fate is negligible, so his private preference plays no direct role in his choice of whether to side publicly with the opposition. His public preference depends on a tradeoff between two payoffs, one external and the other internal.

The external payoff to siding with the opposition varies positively with S . The larger S , the smaller the individual opponent's risk of being persecuted for his outspokenness, and the fewer hostile supporters of the government he has to face. The latter feature reflects the fact that government supporters, even those privately sympathetic to the opposition, participate in the persecution of dissidents, as part of their personal efforts to establish convincing pro-government credentials.

*Associate Professor of Economics, University of Southern California, Los Angeles, CA 90089-0253. I am grateful to the National Science Foundation for support. Most of my research was conducted during a sabbatical, financed partly by the National Endowment for the Humanities, at the Institute for Advanced Study in Princeton.

The internal payoff is rooted in the psychological cost of preference falsification: the suppression of one's wants generates lasting discomfort, the more so the greater the lie. Specifically, person i 's internal payoff to supporting the opposition varies positively with his private preference, x^i . The higher x^i , the costlier he finds it to suppress his antigovernment feelings. An individual's private preference thus plays an indirect role in his choice of a public preference; as a determinant of his internal payoff to supporting the opposition.

Thus i 's public preference depends on S and x^i . As S grows, with x^i constant, there comes a point where the external cost of joining the opposition is outweighed by the internal cost of self-suppression. This switching point is i 's *revolutionary threshold*, T^i . Note that if x^i should rise, T^i will fall.

People with different private preferences and psychological constitutions may differ in their revolutionary thresholds. Imagine a ten-person society featuring the *threshold sequence* $A = \{0, 20, 20, 30, 40, 50, 60, 70, 80, 100\}$. Person 1 ($T^1 = 0$) supports the opposition regardless of its size, just as person 10 ($T^{10} = 100$) always supports the government. The remaining eight people's public preferences are sensitive to S . Initially, the opposition consists of a single person, or 10 percent of the population, so $S = 10$. Because the nine others have thresholds above 10, this S is self-sustaining.

This equilibrium happens to be vulnerable to a minor change in A . Suppose that person 2 has an unpleasant experience with the government, which exacerbates her alienation from the regime. The consequent rise in x^2 lowers T^2 from 20 to 10. Since $S = 10$, person 2 joins the opposition, moving S to 20. This new S is self-augmenting, as it drives person 3 into the opposition. The S of 30 then triggers a fourth defection, and in this manner S feeds on itself until it reaches 90—a new equilibrium. A slight shift in one individual's threshold has thus generated a *revolutionary bandwagon*, an explosive growth in public opposition.

Now consider the sequence $B = \{0, 20, 30, 30, 40, 50, 60, 70, 80, 100\}$, which

differs from A only in its third element: 30 as opposed to 20. As in the previous illustration, let T^2 fall from 20 to 10. Once again, the preexisting equilibrium becomes unsustainable, and S rises to 20. But the opposition's growth stops there, for the new S is self-sustaining. We see that a minor variation in thresholds may alter drastically the effect of a given perturbation.

Neither private preferences nor the corresponding thresholds are common knowledge. So a society can come to the brink of revolution without anyone knowing this—not even those with the power to unleash it, like person 2 in A .

For any number of reasons the threshold sequence may shift dramatically in favor of the opposition. But this will not necessarily trigger a revolution. In the sequence $C = \{0, 20, 20, 20, 20, 20, 20, 20, 60, 100\}$, the average threshold is as low as 30, possibly because in private most people sympathize with the opposition. Yet $S = 10$ remains an equilibrium.

When a revolutionary bandwagon does take off, long-repressed grievances burst to the surface. In addition, people who were relatively content embrace the new regime, attributing their former public preferences to fear of persecution. Reconsider A , recalling that a 10-unit fall in T^2 drives S from 10 to 90. The last person to jump on the bandwagon has a threshold of $T^9 = 80$, a reflection of her great sympathy for the government. Accordingly, she does not switch until the opposition's victory is guaranteed. Having made the switch, she has every reason to feign a longstanding antipathy to the old regime. In doing so, she makes it seem as though the old regime enjoyed even less genuine support than it actually did. This illusion is rooted in the very factor responsible for making the revolution a surprise: preference falsification. Its effect is to make it even less comprehensible why the revolution was unforeseen.

The outlined theory (for details, see my 1989, 1990b papers) unites social evolution and revolution, continuous and discontinuous change, in a single model. Private political preferences and the corresponding thresholds may shift gradually over a long

period during which public opposition is stable. When the cumulative change has established a *latent bandwagon*, a minor event may precipitate a sharp jump in public opposition.

II. The Revolution of 1989

Given communism's failures, the existence of East European dissent is easily understood. Less comprehensible is the rarity of dissent—prior, that is, to 1989. For decades, East Europeans displayed a remarkable capacity to put up with tyranny and inefficiency.

This subservience is attributable partly to punishments the communist establishment imposed on nonconformists. Yet official repression is only one factor in the durability of communism. It met with the approval of disillusioned citizens and relied crucially on their complicity. People with every reason to despise the status quo applauded politicians they mistrusted, joined organizations whose mission they opposed, and signed defamatory letters against dissidents they admired, among other manifestations of consent and accommodation.

In a famous essay, Václav Havel (1979) speaks of a greengrocer who places in his window the slogan "Workers of the World, Unite!" Why does he do this, Havel wonders, "Is he genuinely enthusiastic about the idea of unity among the workers of the world? Is his enthusiasm so great that he feels an irrepressible impulse to acquaint the public with his ideals? Has he really given more than a moment's thought to how such a unification might occur and what it would mean?" (p. 27). No, the greengrocer does not mean to express his real opinion about anything. He displays the slogan simply for the right to be left alone.

The greengrocer's prudence has an unintended consequence: it reinforces the perception of a society united behind the Party. It thus becomes a factor in other people's willingness to continue doing and saying the things expected of *them*.

Later in the same essay, "something in our greengrocer snaps" and he makes "an attempt to *live within the truth*" (p. 39). As a

consequence, he is transferred to the warehouse at reduced pay, and his hopes for a holiday in Bulgaria evaporate. Also, his peers take to harassing him—not out of inner conviction but to avoid being persecuted themselves.

This brilliant vignette suggests that the regimes of Eastern Europe were substantially more vulnerable than the quiescence of their populations made them seem. Millions were prepared to turn against communist rule if ever this became safe to do.

What lowered the level of fear sufficiently to get the revolution underway? With the benefit of hindsight it appears that Gorbachev's reforms in the Soviet Union played a key role. In Eastern Europe these kindled hopes of greater independence and meaningful social change. But why did we not foresee where they would lead?

An examination of the news media before the revolution shows that arguments in the air pointed to the unlikelihood of fundamental change. Even if Gorbachev wanted to liberate Eastern Europe, it was not clear that he could. Surely, Soviet conservatives would insist on retaining their country's security belt. Moreover, tensions within the Soviet Union were sowing the seeds of a conservative coup. Some observers expected Gorbachev to survive, but only by reversing course and becoming increasingly repressive.

For all this pessimism, Gorbachev's policies did fuel expectations of a freer Eastern Europe, reducing the perceived risk of dissent. In terms of our model, they shifted the thresholds of East Europeans increasingly in favor of revolt, making it ever easier to spark an explosion. But obvious as this was, no one could see that public sentiment would shift so soon and so massively.

Pinpointing the specific event that pushed the bandwagon over the hill is akin to identifying the cough responsible for a flu epidemic. There were several turning points, any one of which might have altered history. One came when East German officials cancelled Party leader Honecker's order to fire on demonstrators in Leipzig. The demonstration's peacefulness made many more East Germans sense that change was im-

minent. Another turning point came with Gorbachev's remark that his country had no right to interfere in the affairs of its neighbors. At the time, some East European leaders were contemplating the use of force, and this statement may well have been a major factor in their exercising restraint.

When the greengrocers decide that they have had enough, Havel had predicted, East European communism will collapse like a house of cards. So it turned out: when the masses took to the street, support for the status quo just vanished. In one country after another a few thousand people stood up in defiance, joining long-persecuted activists. In so doing they encouraged additional citizens to drop their masks, which then impelled more onlookers to jump in. Before long, fear changed sides: where people had been afraid to oppose the regime, they came to fear being caught defending it. Party members rushed to burn their cards, claiming they were always reformists at heart. Top officials began sensing that they might face retribution for resistance. They hastened to accept the opposition's demands, only to be confronted with bolder ones (for a chronicle of events, see Timothy Garton Ash, 1990).

The East European Revolution has been billed as the triumph of truth over lying. This designation conveys the end of feigned support for communism, but it conceals the continuation of preference falsification. Lying has not ceased but changed character. Now it provides cover to East Europeans afraid to admit their yearnings for the old order.

III. Is It Surprising that We Were Surprised?

It is tempting to attribute our amazement at the events of 1989 to the inadequacy of our theories concerning political stability. Our most popular theories of revolution certainly left us ill-prepared for the suddenness with which public sentiment turned. For instance, Theda Skocpol's (1979) "structuralist" theory, which shows how changes in international relations can produce social uprisings, does not explain the involved discontinuities. A solid under-

standing of the interdependencies among individual public preferences (whose significance Skocpol explicitly rejects) would doubtless have prepared us better for an East European explosion.

Yet, once again, these interdependencies are largely hidden from view. And for reasons explained above, the knowledge that preference falsification is pervasive does not suffice to establish that a revolution is imminent. We can sense that multitudes are seething with unarticulated discontent without knowing what it would take to turn the possibility of revolt into reality. In principle, of course, we can develop techniques for uncovering the relevant interdependencies. But for all practical purposes we lack the means to find and process all the requisite information. Also, there are irremovable political obstacles to the correct interpretation of whatever information is produced. In view of all this, it is safe to say that no theorizing could have prepared us adequately for 1989.

I ought to point out that this is not the first time a major uprising has come as a surprise. The French Revolution of 1789, the Russian Revolution of February 1917, and the Iranian Revolution of 1979-80 are among the successful revolutions that stunned their leaders, participants, victims, and observers. The failed ones include the Hungarian uprising of 1956 and the Prague Spring of 1968. In all these cases, preference falsification was a prime factor in the suddenness with which public sentiment shifted—and in the cases of failure, shifted back.

Because preference falsification afflicts politics in every society, though in varying forms and degrees, we are likely to be surprised again and again. But obstacles to predicting particular revolutions do not rule out the production of useful general insights into the *process* of revolution. Even if we cannot predict the time and place of the next big uprising, we can prepare ourselves mentally for the mass mobilization that will bring it about. Equally important, we can understand why it may surprise us. There are many spheres of knowledge where useful general theories foreclose reliable

predictions of specific outcomes. The Darwinian theory of biological evolution illuminates the process whereby species evolve, but without enabling us to specify the future evolution of the swordfish.

The theory of biological evolution and the present argument have a common virtue: each reveals the source of its predictive limitations. In the case at hand, the source is *imperfectly observable nonlinearity*. In ways that we cannot hope to grasp fully, public preferences depend on their determinants nonlinearly. This is why an intrinsically insignificant event may generate a massive rise in public dissent.

The notion that small events may unleash huge forces goes against much of twentieth-century social thought, with its emphasis on linearity and thus continuity and gradualism. So does my suggestion of inescapable unpredictability. Lest this be considered offensive to the scientific spirit, I should note that establishing the limits of knowledge is itself a contribution to the pool of useful knowledge. As Friedrich von Hayek (1974) reminds us, it is also necessary for charting a realistic scientific agenda.

REFERENCES

- Ash, Timothy Garton, *The Magic Lantern: The Revolution of '89 Witnessed in Warsaw, Budapest, Berlin and Prague*, New York: Random House, 1990.
- Fischhoff, Baruch and Beyth, Ruth, "I Knew It Would Happen"—Remembered Probabilities of Once-Future Things," *Organizational Behavior and Human Performance*, February 1975, 13, 1–16.
- Havel, Václav, "The Power of the Powerless," in his et al., *The Power of the Powerless: Citizens against the State in Central-Eastern Europe*, (1979) Paul Wilson, trans., Armonk: M. E. Sharpe, 1985.
- Hayek, Friedrich August von, "The Pretence of Knowledge" (1974), *American Economic Review*, December 1989, 79, 3–7.
- Kuran, Timur, "Sparks and Prairie Fires: A Theory of Unanticipated Political Revolution," *Public Choice*, April 1989, 61, 41–74.
- , (1990a) "Private and Public Preferences," *Economics and Philosophy*, April 1990, 6, 1–26.
- , (1990b) "Now Out of Never: The Element of Surprise in the East European Revolution of 1989," mimeo., University of Southern California, October 1990.
- Skocpol, Theda, *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*, Cambridge: Cambridge University Press, 1979.
- Institut für Demoskopie Allensbach, "East German Survey," February 17–March 15, 1990, Archive No. 4195 GEW.

Nations and States: Mergers and Acquisitions; Dissolutions and Divorce

By DONALD WITTMAN*

Many of the most important events in history are associated with the consolidation or dissolution of nations. Alexander the Great founded an empire in a very short time, but its breakup was even quicker. The decline and fall of the Roman Empire has served as a symbol of the temporality of nations. In the more recent past, Hong Kong, Biafra, Eritria, and Northern Ireland are names commonly associated with issues of sovereignty. Within the past year, newspaper headlines have been devoted to the mergers and dissolutions of nations. "Iraq annexes Kuwait," "East and West Germany unite," "Quebec demands greater sovereignty," "Solvenia declares independence" and "The European Economic Community institutes legal reforms" are a small subset of recent examples.

In this article, I provide insight into these historical events by developing an analytic framework for understanding the unification and dissolution of nation states. The basic insight is that wealth maximization (broadly defined) is the underlying fundamental determining the size of nations. I briefly outline some of the efficiency explanations for the size of the polity, then provide a detailed historical example.

I. Efficiency Explanations

As my title suggests, much of our insight regarding states comes from the analogous situation regarding firms and marriages. Firms acquire other firms if the value of the joint firm is greater than the total value of the firms as separate entities. Assets are divested when the assets are of greater value

in the hands of others. Countries should merge when the economic value is greater for a unified country than as separate sovereign states. In theory, economic value is unequivocal if those who want union can bribe those who do not want union, but those who do not want union cannot bribe those who want union to change their minds.¹

Economies of scale and scope are two basic explanations for the size of nations. The costs of administration and policy coordination are unlikely to be proportional to the size of the polity. In warfare especially, there are often great gains in military strength when two small states merge. Firms may create cartels to lessen competition among themselves thereby gaining an advantage on either their suppliers or demanders. In a similar way, countries may create alliances or even merge in order to reduce competition among themselves and gain an advantage over others not in the alliance. At some point, diseconomies of scale arise. The existence of many nations suggests that economies of scale and scope are exhausted before unitary political control of the world is reached.

Different political systems have different economies of scale. Participatory democracy puts severe constraints on the size of the polity. In communist systems, the political apparatus is intricately intertwined with the economic sector. Other things being equal, the optimal size of communist nations is likely to be larger than for capitalist nations since the political scope of the former is so

*University of California, Santa Cruz, CA 95064. I thank the discussant, Martin McGuire, and my colleagues at UCSC for their helpful comments. This article is excerpted from a more extensive working paper with the same title.

¹This model differs from the seminal work of David Friedman (1977) who views each nation as Leviathan. For example, in this model, the shape of the country allows for minimal migration and hence maximal taxation of labor. His model does not consider the desire by the populace to minimize the ability of a country to undertake such extortion.

much more extensive.² An international regime of free trade among nations allows for smaller-scale political units since economic production is not limited to domestic markets. Administrative technology also plays a role. Préliterate societies tend to be less extensive than those that can keep records.

Synergy arises when the merger of the two unique nations produces more than the sum of the parts. Synergy is maximized when the set of merged states creates greater wealth than any other combination of states. It seems plausible that the combined energies of Madison, Hamilton, and Jay in creating *The Federalist Papers* and the ensuing federation was greater than if each had spent his energies in developing his own separate state (Virginia, New York, and New York). Africa is the great natural experiment where one can observe the various combinations of colonial powers with indigenous tribes and their residual effect on the polity after independence was achieved. At one time, Lebanon was an example of synergy. The combination of Christian and Moslem populations created a cosmopolitan country, attracting tourists from the stricter neighboring Moslem countries and serving as a financial center for the Mideast.³ Negative synergy arises when the combination creates disfunction.

In the economic sector, sometimes the best way for a firm to gain access to the *unique* skills associated with another firm is to become part of the more successful firm. Thus one firm may take over another because the firm's CEO and administrative apparatus will create more wealth for the targeted firm than its existing administration.

Similar phenomena occur within the political sector. Consider the recent unifica-

tion of Germany. The economic advantages to East Germany of uniting with West Germany are quite obvious. There are also great political advantages. Rather than creating a democratic society out of whole cloth, the East Germans gained liberal democratic and relatively noncorrupt institutions at low cost by becoming part of united Germany.⁴ A closer analogy to a stock takeover is the secession of Charles II of Spain in 1700. England, France, and the United Provinces were scheming to divide Spain. Rather than accept such a division, the king offered all of his realm as an indivisible bequest to the strongest of the claimants, France, but only if the integrity of the country and monarchy remained; if Phillip, the grandson of Louis XIV, refused to abide by this, then the crown would be bequeathed to Archduke Charles of Austria on the same terms. The installation of Phillip V of Spain brought fresh energy and new administrative methods to the decaying Spanish kingdom.

The merger of states reduces interstate transaction costs but increases intrastate transaction costs. The impetus to economize on transaction costs affects the number and size of nations. The appropriate forum (within a merged nation or across two sovereign nations) for resolving conflict depends on the political institutions available in each sphere and the nature of the potential conflict. The existence of the World Court, the willingness of nations to forego war as a means of settling disputes, and other methods of reducing interstate transaction costs reduce the optimal nation size. More generally, different national and transnational political institutions have different transaction costs and thus the comparative advantage of size depends on which political institutions are in place.⁵

² Communist systems often have inappropriate pricing mechanisms making international trade difficult. The resulting need for autarky creates a greater demand for a larger country.

³ Unfortunately, the rigid power-sharing structure and weak central government made Lebanon unable to cope with the arrival of the PLO, shifting demographics, and interference from its neighbors.

⁴ What were the advantages to W. Germany in joining with E. Germany? In the short run, there are great costs. But in the long run, the free movement of capital and labor will greatly increase the wealth of both countries.

⁵ I will not consider the endogeneity of political structure. I want to go beyond the standard public finance explanation for jurisdictions—that the extent of the jurisdiction covers the extent of the externality.

II. Historical Example

The ideas developed above can provide insight into historical events. My example of France and Poland in the nineteenth century is chosen for illustrative purposes only; it demonstrates the stance one takes in explaining history. A more thorough discussion would involve a separate article.

While not denying the role of the military in shaping the political map, my analysis places military power in a secondary position. First, military might and even military victory need not lead to territorial changes. For example, after Napoleon's defeat and his banishment to Elba, the Congress of Vienna extracted virtually no European territorial concessions from France although the allied armies had ended the war at the gates of Paris. Even after Napoleon's return and subsequent defeat at Waterloo, territorial concessions were minimal (the boundaries of France remained virtually unchanged from those at the start of the French Revolution); instead, more severe indemnities were to be paid by the vanquished nation. The end of the Franco-German War in 1871 found Paris surrounded, yet territorial concessions were limited to Alsace and Lorraine and control over the Papal territory. Instead of more territory, Germany received 200 million pounds sterling. A trade had been made. If they were rational, there was no other bargain (for example, France giving up more territory, but paying less indemnity) that would have made both sides better off.

Second, military might is often a result of other variables. These variables *may* be the underlying explanation for territorial change. The ability to raise large armies and finance a war depends on economic strength and a very effective political apparatus. Political skills needed to wage a war are positively correlated (but not perfectly) with those skills necessary for ruling a nation.

Napoleon is often viewed as a great military strategist, but he was also an effective administrator and statesman. Furthermore, his success was built on the many reforms of the French Revolution (including standardized taxes and centralized power).

Expansion of one country at the expense of another may be due to the more effective political structure and competent staffing of the expanding country (which may be reflected in their relative military capabilities). The same holds true for the reverse: a successful war of independence may result from the center being too politically weak to either rule the break-away state effectively or create a powerful military force.

A comparison of nineteenth century Poland and France is insightful. The flatness of Poland has always been used as a reason for its easy subjugation, but the partition and complete disappearance of Poland in the latter part of the nineteenth century was also due to its internal weaknesses. Polish nobility had been a very powerful force in maintaining the status quo; anyone could veto legislation and dissolve the assembly. The central government had neither substantial revenues, nor a well-equipped standing army, nor a bureaucracy to do the work of a state. Poland was in need of a more effective political apparatus. It got one by being gobbled up by its neighbors. A duchy remained under the control of Russia, but it did not maintain the name of Poland.

In contrast, France maintained most of her territorial integrity, despite having much of her territory occupied in three separate wars. An important difference between the two countries was that, with the exception of those times of internal strife, France had an effective national apparatus while Poland did not. To continue with the takeover metaphor, France is an example of green-mail while Poland is an example of a hostile takeover.

III. Summary

It is hard to condense the history and theory of the size of nations into a short paper. Even restricting the viewpoint to that

Since "externalities" can be dealt with across jurisdictions in a Coasian bargain, externalities alone cannot explain the size of the political entity.

of economic theory provides many disparate components. Therefore, I merely summarize some of the major stands of my argument.

I have argued that the role of military power in explaining the takeover (or breakup) of nations, while not insignificant, has been overemphasized—it is often an effect rather than a cause. At times, military strength may reflect underlying political power. A successful war of independence may occur because the center was too weak politically to either rule the break-away state effectively or create a powerful military apparatus. At other times, a more powerful military force may reflect the greater value to the country from occupying certain territory; that is, willingness to pay for military action partially depends on the gain from such activity. The biggest argument against the equality between military power and size of nations is that the conquering country has often given up captured territory for monetary and other indemnities.

When commitments are credible, then the size of nations tend toward the optimal (see my 1991 paper). Two nations would join together (separate) if the economies of scale and scope and the synergy produced by their union created greater (smaller) benefits than costs. Militarism still might be used by the strong to exploit the weak, but the tribute paid by the weak would not necessarily be territorial.

The merger of states reduces interstate transaction costs but increases intrastate

transaction costs. As international organization improves (the creation of the European Economic Community) and international transaction costs are reduced (most importantly, the threat of war is decreased and barriers to international trade are reduced), then the size of nations will be reduced. This is occurring in the Soviet Union and, to a lesser degree, even in Great Britain. As European integration increases, British integration is likely to decrease, with Scotland, for example, taking on a more independent role within Great Britain. Similar forces are at work in China. Power is shifting from the center to the provinces.

I have purposely omitted two of the standard arguments for nations—nationalism and ideology. Economists do not use such devices to explain firm conglomeration and divestiture and they should not use them to explain nation size either. The analysis need not be restricted to the political dimensions of nations. Many of the insights provided here can be applied to cities, religions, tribes, and clans.

REFERENCES

- Friedman, David, "A Theory of the Size and Shape of Nations," *Journal of Political Economy*, February 1977, 85, 59–77.
- Wittman, Donald, "War and Mergers of Sovereign States: The Role of Credible Commitments," University of California-Santa Cruz Working Paper, 1991.

The Technology of Conflict as an Economic Activity

By JACK HIRSHLEIFER*

People can satisfy their desires in two main ways: by *production* (for self-use, or for mutually beneficial trade with other parties), or else by *conflict* (i.e., by actual or threatened theft, robbery, confiscation, or litigation). Despite its evident importance, only recently has a systematic economics of struggle and conflict begun to emerge.¹

I. Technology and Conflict

Conflict, as opposed to mere failure of cooperation, comes about when one or more parties calls upon a special technology. To wit, a technology whereby some or all contenders for resources incur costs in an attempt to weaken or disable competitors. This definition is broad enough to encompass not only war but also strikes and lockouts, lawsuits, sibling rivalries within families, and redistributive politics. But, for concreteness here, I mainly use the metaphorical language of military combat.

The costs of conflict as an economic activity can include: 1) foregone opportunities, as when guns are produced rather than butter; 2) attrition of the resources actually devoted to combat, for example, military casualties; 3) collateral damage to productive resources. The prospect of collateral damage, intentional or unintentional, reduces the profitability of conflict. In fact, the retaliatory threat of collateral damage is, according to modern deterrence theory, the key to peace in a nuclear age. As a nonmilitary example, subsidy seeking is held somewhat in check to the extent that prospective beneficiaries recognize they will be incurring some share of the collateral

damage in the form of deadweight loss to the economy (Gary Becker, 1983). And scorched-earth tactics (imposing or threatening to impose collateral damage on oneself) may play a role not only in warfare but in corporate takeover struggles.

An operations research literature exists that looks into what might be called the industrial engineering of combat: for example, the vulnerability of submarines to depth bombs, the effectiveness of anti-aircraft guns placed on merchant ships, the tradeoff between training air crews versus actual mission time (Philip Morse and George Kimball, 1951). Of more interest to economists is the *macrotechnology* of conflict: how generalized resources devoted to struggle generate outputs in the form of gains and losses to each side.

Such an economic analysis of conflict may help explain not only the prevalence of war and peace, but also the sizes and shapes of nations, the distribution of power among social classes, the viability of two-party vs. multiparty political systems, strikes and lockouts in industry, sibling rivalry, why there are more lawsuits nowadays, and who wins in the battle of the sexes. On this vast range of topics I can offer here only some preliminary ideas, focusing upon how outcomes are constrained by the technology of conflict.

II. Conflict Models

Modeling of conflict can be said to begin with Frederick Lanchester (1916).² In his best-known formulation, letting F_1 and F_2 be force sizes at any moment, and α_1 and α_2 be fighting efficiency parameters, the attrition rates of the two forces can be repre-

*Department of Economics, University of California, Los Angeles, CA 90024.

¹I cite here only three pioneering works: Kenneth Boulding (1962), Thomas Schelling (1960), and Gordon Tullock (1974).

²Robert Samz (1970) surveys a number of precursors of Lanchester.

sented by

$$(1) \quad dF_1/dt = -\alpha_2 F_2; \quad dF_2/dt = -\alpha_1 F_1.$$

Solving the differential equations indicates that the relative strengths vary with the squares of the force sizes. Whichever side is initially stronger inevitably grows stronger still, until the weaker is extinguished. This is probably too extreme, but even if it were valid, we would still have to consider both the costs and benefits of military victory.

In a simple version of such a model, each side $i = 1, 2$ seeks to maximize its steady-state income I_i . We can suppose that: (a) each side divides its endowed resources R_i between productive effort E_i and fighting effort F_i ; (b) aggregate income I is generated by the two sides' productive efforts; and (c) the respective shares p_i depend upon the fighting efforts. Thus:

$$(2a) \quad E_i + F_i = R_i$$

$$(2b) \quad I = \Omega(E_1, E_2)$$

$$(2c) \quad p_1 = \Gamma(F_1, F_2); \quad p_2 = 1 - p_1$$

$$(2d) \quad I_i = p_i I.$$

To actually solve this general equilibrium system, we would also have to specify a solution concept: Cournot, Stackelberg, or whatever.

The specification of the *productive technology* in (2b) involves some drastic simplifying assumptions, among them that: (i) all income I is jointly produced, and falls into a common pool available for capture; (ii) conflict activity does not damage the resources R_i or the productive process that generates income. In effect, the only costs of conflict recognized are the opportunity losses of devoting resources to fighting effort F_i rather than to productive effort E_i . The *conflict technology* Γ of equation (2c), which might be termed the Contest Success Function (CSF), is oversimplified, among other respects, by making no distinction between offense and defense. Or, looking at this in another way, the model does not

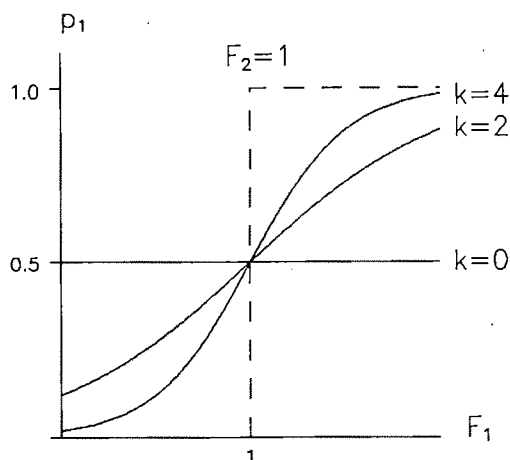


FIGURE 1. CONTEST SUCCESS FUNCTION AS THE MASS EFFECT PARAMETER k VARIES

provide any special advantage for status quo possession.

An attractive form for the CSF is the logistic function:

$$(3) \quad p_1/p_2 = \exp\{kF_1\}/\exp\{kF_2\}$$

$$\text{or} \quad p_1 = 1/[1 + \exp\{k(F_2 - F_1)\}].$$

Here k is a "mass effect parameter" scaling the decisiveness of fighting effort disparities. As plotted in Figure 1, the horizontal axis represents side No. 1's force F_1 , where by assumption the opposing force is fixed at $F_2 = 1$. (The efficiency parameters α_i are assumed equal to unity.) On the vertical axis is plotted p_1 , No. 1's fractional share of income (or, on an alternative interpretation, No. 1's *probability of winning*).

The logistic function squares with what might be called the first "stylized fact" of warfare: the tremendous advantage of being even just a little stronger than one's opponent. In more familiar economic terminology, we would expect to observe *increasing marginal returns* to force size F_1 so long as $F_1 < F_2$, with *decreasing marginal returns* holding thereafter—as is broadly consistent with military experience (T. N. Dupuy, 1987). Note also that p_1 remains positive even when zero fighting effort is invested, a feature consistent with a second stylized

fact: that there can be a peace either of subjugation (one side finds its optimal to choose zero fighting effort) or else of balanced settlement (both sides find it optimal to do so).³ Another implication of the logistic CSF is that *proportionately growing wealth* on the two sides works to the net advantage of the wealthier party. Thus, over time, smaller contenders may tend to be subjugated by larger ones, which perhaps helps explain a third stylized fact: the tendency toward fewer but larger states over historical time.

Underlying the logistic form is the premise that the numerical difference $F_1 - F_2$ between the fighting efforts determines conflict success. An alternative approach would be, as in the Lanchester model, to make success a function of the ratio F_1/F_2 . The ratio form turns out to be inconsistent, or only imperfectly consistent, with the stylized facts listed above. In particular, peace is impossible since a side unable or unwilling to fight loses everything. One interpretation is that the ratio form applies under "ideal" conditions such as an undifferentiated battlefield, full information, and unflagging effectiveness. But where what Clausewitz called "friction" plays a role (if there are sanctuaries and refuges, if information is imperfect, or where fatigue limits what a victor can do), the difference form better describes the outcome.

Foregoing possible generalization, for lack of space I turn instead to another question: the competition between the technology of conflict and the technology of peaceful production and exchange.

III. Technology of Conflict vs. Technology of Production and Exchange

For concreteness, suppose the production function Ω is such that

$$(4) \quad I = \Omega(E_1, E_2) = A(E_1^{1/s} + E_2^{1/s})^s.$$

Here A is a measure of the total productiv-

ity of the "factors" E_1 and E_2 , while s is an index of the complementarity between them. For given inputs, as s rises, income I will increase as well, and the marginal products $\partial I/\partial E_1$ and $\partial I/\partial E_2$ will also both rise.

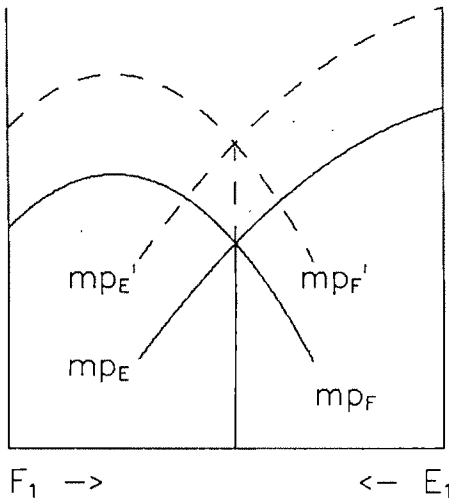
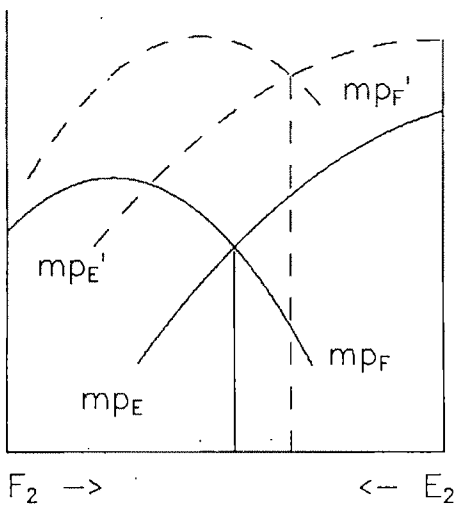
In Figure 2a, given the E_2, F_2 choice of the other side, for contender No. 1 the mp_F curve shows the marginal product of fighting effort while mp_E represents the marginal product of productive effort. No. 1's best-response optimum is at the intersection. However, the curves will themselves shift in response to the opponent's choice of F_2 vs. E_2 , and they will also be sensitive to exogenous factors like changes in preferences, in wealth endowments, or in the Ω and Γ functions.

Suppose the complementarity index s between the parties' productive efforts were to rise. Then, as indicated in Figure 2a, No. 1's mp_E curve shifts upward. It might therefore seem that increased complementarity implies rationally devoting more effort to production and less to conflict. However, the increase in I puts more income at stake, so that fighting effort also tends to have a higher return. Thus the mp_F curve also shifts upward, so that the net effect remains unclear. This consideration exposes the fatal flaw in assertions that the growing interdependence among nations is making war obsolete.⁴ That same growing interdependence also typically raises the stakes of conflict—each side has more to gain from fighting and more to lose by failing to fight.

In Figure 2b, taking contender No. 2 as protagonist, we ask what happens when opponent No. 1 becomes richer than before. Normally, No. 2 can expect that an enriched No. 1 will be choosing higher levels both of productive effort and of fighting effort—larger E_1 and larger F_1 . Given positive complementarity, the larger E_1 will make the mp_E curve for No. 2 shift upward, so there

³See my 1988 article; see also Stergios Skaperdas (1990).

⁴Just before World War I, an immensely popular book by Norman Angell (1911) argued that the growing economic absurdity of modern war meant that the major powers of Europe would no longer find it advantageous to fight one another—or, at least, to persist in such a conflict for very long. Similar arguments are still heard today.

Marginal
productFIGURE 2a. EFFECT OF INCREASED
COMPLEMENTARITYMarginal
productFIGURE 2b. WHEN NO. 1 GET RICHER,
NO. 2 FIGHTS HARDER

is a force tending to induce No. 2 to revise his choice in the direction of productive effort E_2 . But there are two more powerful considerations working the other way: (i) aggregate income being higher, as before there is more of a prize to be fought over,

and (ii) the opponent's F_1 being larger means that No. 2 must invest more fighting effort just to retain what he already had. So a less well-endowed side will find it optimal to devote relatively more effort to fighting: "No. 2 tries harder."

IV. Some Implications of Conflict Technology

Individuals and groups can compete by employing the technology of conflict as an alternative to the technologies of production and exchange. To summarize only a tiny sample of the implications: *As regards peace or war*: Peace is more likely to the extent that the decisiveness of conflict is low or (though I did not have time to develop these points) if the stakes are small or if the technology favors the defense. More surprisingly, perhaps, increased productive complementarity between the parties does not systematically favor peace. *As regards relative success*: Evidently, a favorable force disparity $F_1 - F_2$ promotes success, thus conferring an advantage upon the better-endowed side. Also, greater decisiveness of conflict increases the benefit of such a disparity. However, the poorer side is generally motivated to invest relatively more heavily in fighting effort. So conflict can be an income-equalizing process. In fact, newly impoverished social groups do typically become more bellicose, newly enriched groups more pacific and accommodating.

REFERENCES

- Angell, Norman, *The Great Illusion*, New York: G. P. Putnam, 1911.
- Becker, Gary S., "A Theory of Competition Among Pressure Groups for Political Influence," *Quarterly Journal of Economics*, Aug. 1983, 98, 371-400.
- Boulding, Kenneth E., *Conflict and Defense*, 1962; republished Lanham: University Press of America, 1988.
- Dupuy, T. N., *Understanding War*, New York: Paragon House, 1987.
- Hirshleifer, J., "The Analytics of Continuing Conflict," *Synthese*, August 1988, 76, 201-33.
- Lanchester, Frederick William, *Aircraft in*

- Warfare: The Dawn of the Fourth Arm*, London: Constable, 1916; Extract in J. R. Newman, ed., *The World of Mathematics*, Vol. 4, New York: Simon and Schuster, 1956, 2138-57.
- Morse, Philip M. and Kimball, George E.**, *Methods of Operations Research*, 1st ed. rev., Cambridge; and New York: Technology Press; Wiley & Sons, 1951.
- Samz, Robert Walter**, "Toward a Science of War through some Mathematical Concepts of Macrocombat," unpublished doctoral dissertation, Arizona State University, 1970.
- Schelling, Thomas C.**, *The Strategy of Conflict*, Cambridge: Harvard University Press, 1960.
- Skaperdas, Stergios**, "Cooperation, Conflict and Power in the Absence of Property Rights," University of California-Irvine, May 1990.
- Tullock, Gordon**, *The Social Dilemma*, Blacksburg: Center for the Study of Public Choice, VPISU Press, 1974.

GREENHOUSE WARMING[†]

International Trade in Carbon Emission Rights: A Decomposition Procedure

By ALAN S. MANNE AND RICHARD G. RICHELS*

In recent years, a number of proposals have been advanced for the limitation of carbon emissions. Some have argued that such limits would be costless, but our analysis suggests that there is no free lunch. (See our forthcoming paper.) We have attempted to estimate the costs but *not* the global benefits of slowing down climate change through carbon limitations. All computations were performed in parallel for five geopolitical regions. Except for oil trade, these regions were treated independently—as though there were no opportunity for international trade in carbon rights. For stimulating ideas on the politics and economics of negotiating an agreement on greenhouse gas emission permits, see M. Grubb (1989).

Whatever rule is adopted for the allocation of carbon emission rights, there are likely to be significant interregional differences in the value of these rights. International trade will be needed if economic efficiency is to be achieved. In the absence of such trade, there are likely to be significant distortions in the comparative advantage of individual locations for the production of

tradeable basic materials such as primary metals. These distortions could lead to counterproductive regulations and new forms of nontariff barriers to trade. This paper is intended to quantify the potential for international trade in carbon emission rights.

I. The Decomposition Procedure

Each individual region is viewed as a price taker, and as a possible importer or exporter. Each is coupled to the others through the international price of these rights. Since this is an intertemporal problem, the time path of prices must be determined so as to equilibrate supplies and demands during each period simultaneously. We can no longer formulate the overall problem as five independent nonlinear *optimizations*. Instead, we must solve an *equilibrium* problem in which all choices are integrated through an international market in carbon rights—even though each of the five agents has an independent objective function and resource endowments. For a masterful review of such computations, see H. Scarf (1984).

An equilibrium problem is easy to formulate, but this one suffers from the *curse of dimensionality*. Since each commodity is differentiated between regions and time periods, there are approximately 5000 prices and quantities to be determined. A problem of this size is about five times larger than can be handled by the best algorithms currently available. (See, for example, T. Rutherford, 1989.) As a practical alternative, we have turned to a decomposition procedure.

The decomposition principle was originally applied to linear programming by

[†]*Discussants:* James Broadus, Woods Hole; Geoffrey Heal, Columbia University; James Sweeney, Stanford University.

*Stanford University, Stanford, CA 94305, and Electric Power Research Institute, Palo Alto, CA 94303, respectively. The research was funded by the Electric Power Research Institute (EPRI). The views presented here are solely our own and do not necessarily report the views of EPRI or its members. For research assistance, we are indebted to Diane Erdmann, Lawrence Gallant, and Robert Luenberger. Helpful comments have been provided by Lawrence Goulder, Stephen Peck, Scott Rogers, Thomas Rutherford, John Rowse, and John Stone.

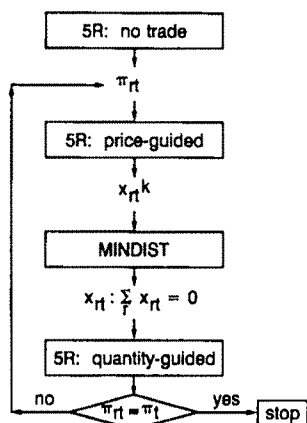


FIGURE 1. DECOMPOSITION PROCEDURE

G. Dantzig and P. Wolfe (1961). These ideas were later adapted to a nonlinear equilibrium model of international trade by A. Mansur and J. Whalley (1982). This paper reports numerical results employing a procedure motivated partly by Dantzig-Wolfe and partly by Mansur-Whalley. There is price-guided decentralization, there are quantity responses from the individual regions, and there are quantity allocations of tradeable goods at the final iteration. Although the method works well here, there is no proof that it will converge under all circumstances. These issues are left as a challenge to subsequent investigators.

Figure 1 provides an outline of the procedure. 5R is an acronym that describes the parallel five region optimizations. These are performed in three modes: no trade in carbon rights, price-guided decentralization, and quantity-guided decentralization. In order to solve the master problem, we employ MINDIST (a minimum distance convex combination of the proposals received from the individual regions).

The only tradeables consist of carbon rights and a numeraire good produced in each region. Let π_{rt} denote the price of carbon rights in region r during period t . In equilibrium, this price will be identical in all regions. That is, $\pi_{rt} = \pi_t$ (for all t).

Let x_{rt} denote the net exports of carbon rights from region r during period t . Note that x_{rt} may be positive, negative, or zero.

In equilibrium, $\sum_r x_{rt} = 0$ (for all t). That is, the international market for carbon rights must clear during each period t . Our decomposition procedure is designed so that the *quantity* equilibrium conditions are met exactly. The *price* equilibrium conditions are met only approximately. If there are significant discrepancies in these prices, one cycles back to the second step of the algorithm and eventually arrives at a more efficient allocation of carbon rights.

II. An Illustrative Example

In this example, there is an international agreement that total carbon emissions be limited to their 1990 level of 5.7 billion tons during each year of the twenty-first century, but the shares in carbon rights change gradually over time. Initially (during the year 2000), carbon rights are distributed between regions in proportion to their 1990 level of emissions. At the end of the planning horizon (the year 2100), carbon rights are distributed in proportion to the 1990 level of population. This proposal is designed to avoid an abrupt change in the status quo, but over the long run it leads to an egalitarian distribution of carbon rights. The 1990 population base is chosen specifically so as to penalize those nations that fail to control their rate of population growth.

According to Figure 1, the first step in the decomposition procedure is to determine the carbon prices that would prevail within each individual region in the absence of trade. That is, we set $x_{rt} = 0$, and solve 5R for the prices π_{rt} . Figure 2 contains a typical set of numerical results for USA (United States), OECD (other OECD); SU-EE (Soviet Union and Eastern Europe); China and ROW (rest of world). Despite the differences from one region to another, there is a general pattern to these price trajectories. During the initial years, it is optimal to delay the use of carbon rights. Prices then rise at the same rate as the marginal productivity of capital in each region—about 5 percent annually. During the later years, carbon prices stabilize at \$250/ton. This level is determined by the cost difference between a carbon-free and a car-

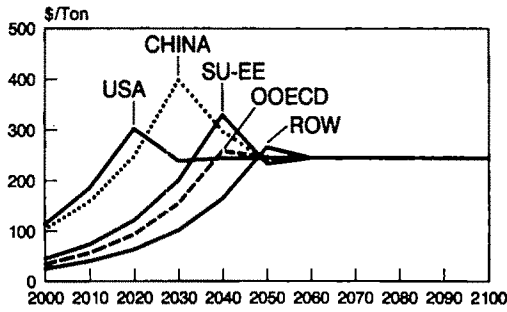


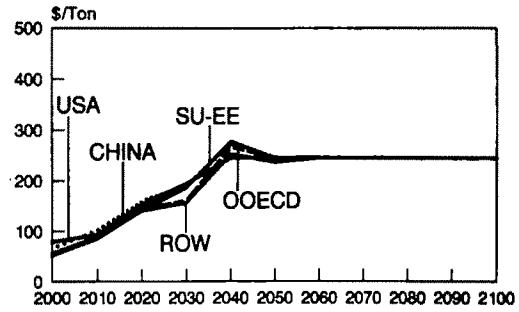
FIGURE 2. CARBON PRICES—NO TRADE

bon-based backstop and by the carbon coefficient of the backstop. There is also an intermediate period during which the new technologies require time for market penetration, and the carbon tax therefore overshoots the backstop level.

Figure 2 provides an important insight. It takes no more than a back-of-envelope calculation to determine the carbon tax during the backstop phase. If all regions have identical energy supply options at that time, there is no motivation for trade. Each region can be self-sufficient in carbon rights. Alternatively, if there are systematic differences between regions, the carbon tax during the backstop phase will be determined by the least-cost region, and it will export carbon rights to the others. There is no need to construct an intertemporal equilibrium model in order to analyze international trade in carbon rights during the backstop phase.

An intertemporal model is useful only during the transition period—when prices are first rising and then falling toward the backstop level. Moreover, an intertemporal model is needed in order to indicate the beginning date for the backstop phase. A major simplification is suggested by Figure 2. It suggests that we can truncate the analysis of international trade in carbon rights in the year 2050. Virtually all five regions have reached the backstop phase by that point.

Now return to Figure 1. For each region r , the "no trade" price vectors $\pi_{r,t}$ are internally consistent. That is, these prices do not rise more rapidly than the marginal produc-

FIGURE 3. CARBON PRICES—WITH TRADE
(CYCLE 4)

tivity of capital. We may now use these price vectors in succession to solve 5R for the case of price-guided decentralization, allowing for positive or negative values of the net export variables $x_{r,t}$. Thus we first solve 5R using $\pi_{USA,t}$, the price vector that emerged from solving the USA submodel under the assumption of no trade. Similarly, we solve 5R using the price vectors $\pi_{OECD,t}, \dots, \pi_{ROW,t}$.

In general, we solve 5R for each of K carbon price vectors, and record the regional distribution of net exports as $x_{r,t}^k$ (for $k = 1 \dots K$). In the first cycle, there are only five carbon price vectors, and $K = 5$.

The principal inputs to the "master" problem are the net export records associated with each of the K price vectors. For a given price vector k , the *price-guided* version of 5R determines $\bar{x}_t^k = \sum_r x_{r,t}^k$ = the global level of net carbon exports during period t . In general, the k th price vector will *not* lead to a global equilibrium in net export quantities; that is, $\bar{x}_t^k \neq 0$. The master problem is formulated, however, so that a weighted average of these net export vectors will add up to zero. The criterion for the choice of weights is *minimum weighted distance* from zero net export demands. With the net export allocations $x_{r,t}$, we run the *quantity-guided* version of 5R (the bottom box in Figure 1). This leads to a new set of region-by-region carbon prices $\pi_{r,t}$. If these are sufficiently close to each other, we can terminate. Otherwise, we can use these prices to begin a second cycle of the price-guided version of 5R, etc. By the fourth cycle, we obtain the prices shown on

Figure 3. Note that the decomposition procedure leads to regional prices that match each other quite closely. That is, $\pi_{r,t} \approx \pi_t$. Using GAMS/MINOS, this procedure requires about four hours on a desk-top computer. See A. Brooke et al. (1988). Further details are available in an appendix that can be obtained upon request to the authors.

III. Additional Results

In this scenario, the United States and China would be the major importers of carbon rights. The Soviet Union and Eastern Europe would be an exporter during the initial years and an importer subsequently. The other two regions (OECD and ROW) would be the major exporters.

Trade could lead to a significant volume of international financial transfers. In 2020, the United States would import .3 billion tons of carbon emission rights. Valued at \$150/ton, this would amount to an expenditure of \$45 billion. It would not be easy for political leaders to justify transfers of this magnitude. On grounds of economic efficiency, however, this would be clearly preferable to a detailed regulatory system in which specific carbon conservation guidelines are mandated for each class of end user.

What are the economic costs of carbon restrictions, and what are the benefits from international trade? According to this example, we should not expect dramatic gains from trade. None of the five regions imports or exports more than 5 percent of the total global volume of carbon rights.

Because of year-to-year fluctuations, it is best to express the gains from trade in terms of their effect upon discounted macroeconomic consumption throughout the planning horizon. Consumption losses are defined to be the difference between consumption with and without limits upon carbon emissions. Employing a 5 percent real rate and discounting back to 1990, we obtain the results shown in Figure 4. With no trade, it would cost the United States about \$1700 billion during the twenty-first century to participate in the international carbon limitations program. With trade, the costs

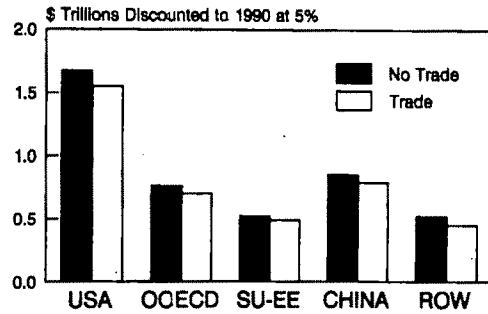


FIGURE 4. CONSUMPTION LOSSES

would be lowered by about \$100 billion. Each of the other regions would also benefit modestly from trade.

Caveat. These numerical estimates should be viewed as illustrative, and are subject to change as we obtain a better understanding of the parameters underlying the model. At this point, our calculations do not allow for indirect effects upon international trade through changes in the production locations of carbon-intensive basic materials such as iron and steel. It is clear that different results will be obtained if we modify any of the input assumptions. This is precisely the purpose of a model. In the debate over the costs of reducing the rate of climate change, our aim is to promote "second-order" agreement—that is, to identify the specific factors that are crucial to an overall assessment of the costs and benefits of limiting carbon emissions.

REFERENCES

- Brooke, A., Kendrick, D. and Meeraus, A., *GAMS: A User's Guide*, Redwood City: Scientific Press, 1988.
- Dantzig, G. and Wolfe, P., "The Decomposition Algorithm for Linear Programs," *Econometrica*, October 1961, 29, 767-78.
- Grubb, M., *The Greenhouse Effect: Negotiating Targets*, London: Royal Institute of International Affairs, 1989.
- Manne, A. and Richels, R., "Global CO₂ Emission Reductions—the Impacts of Rising Energy Costs," *The Energy Journal*, forthcoming 1991.

Mansur, A. and Whalley, J., "A Decomposition Algorithm for General Equilibrium Computation with Application to International Trade Models", *Econometrica*, November 1982, 50, 1547-57.

Rutherford, T., "General Equilibrium Modelling with MPS/GE," Department of

Economics, University of Western Ontario, April 1989.

Scarf, H., "The Computation of Equilibrium Prices," in his and J. Shoven, eds., *Applied General Equilibrium Analysis*, Cambridge: Cambridge University Press, 1984, ch. 1.

Towards a Comprehensive Approach to Global Climate Change Mitigation

By RICHARD D. MORGENSTERN*

Almost a century ago the Swedish scientist Svante Arrhenius coined the term "greenhouse effect." Since then physical scientists have been debating the potential for and implications of a significant change in the earth's climate associated with an increase in anthropogenic emissions of carbon dioxide. Within the past decade, scientists have developed expanded (and presumably improved) models of the global atmosphere, along with a fuller understanding of trace gases other than carbon dioxide that may alter the global heat balance. A number of panels of distinguished scientists, while noting the many uncertainties, have, nonetheless, warned of the risk of serious and, for practical purposes, irreversible global warming associated with projected greenhouse gas (GHG) emissions.

Economists are quite new to the field and are just beginning to frame their issues. Yet even at this early stage, the two economic issues that seem most compelling are the "so what" question and the notion that "maybe the cure is worse than the disease." "So what" refers to the argument that, even if the physical and biological changes on earth are large, maybe the economic consequences of such changes are small. Economists recognize this argument as a call for an aggregate damage function (i.e., benefits of avoided damages). "Maybe the cure is worse than the disease" refers to the allegedly huge costs associated with any attempt to forestall or even significantly delay global warming. Economists recognize this

latter argument as a call for an aggregate mitigation cost function.

To date, William Nordhaus (1990) is the only economist intrepid enough to write about both damage and cost functions associated with climate change. His pioneering works presents a "resource steady-state" model that integrates a discussion of the potential impacts of climate change (i.e., benefits due to avoided damage) with a mitigation cost function for the purpose of determining the optimal level of control of GHGs. His principal conclusion is that beyond continuing the scheduled phase out of CFCs, only a modest reduction of CO₂ (as would be achieved by a carbon tax of perhaps \$5/ton or less) is warranted on an economic basis at this time.

Nordhaus outlines the truly global nature of the climate change issue, but because of data availability focuses his analysis on the United States. This practical approach ignores the very real possibilities that the damages of climate change may be greater and/or the marginal costs of emission reduction lower in developing countries. Both possibilities, of course, imply that the optimal control level of GHGs may be greater for the global economy than the United States alone. Although the present paper is (regrettably) subject to the same data constraints, this is potentially a very important limitation that should be emphasized at the outset. The bulk of this paper represents work in progress.

I. Climate Change Benefit/Cost Framework

The goal of policymaking is to design an efficient climate change mitigation strategy that maximizes (under uncertainty) overall welfare, including measured GDP and the value of nonmarket goods and services (i.e., "Green GNP"). Such a strategy should select an amount of net emissions reduction and/or adaptation at which the marginal

*Director, Office of Policy Analysis, U.S. Environmental Protection Agency, Washington, D.C. 20460. I acknowledge the helpful comments of Jim Broadus, Howard Gruenspecht, and Bruce Schillo. The views expressed herein are my own and do not necessarily represent the views of the EPA or the U.S. government.

benefit of reduced GHG concentrations equals the marginal cost of reducing GHG concentrations or engaging in adaptation. Actual emissions reductions would be achieved at minimum cost by choosing the set of least-cost mitigation options.

Benefits may arise from changing the timing or magnitude of such adverse effects of climate change as lower agricultural yields, impacts of sea-level rise, adverse effects on human health, degradation of recreation and leisure time, and lessened biological diversity. The costs of a mitigation strategy encompass the direct expenditures needed to reduce GHG emissions, to increase the uptake of GHG sinks, or to adapt. Direct mitigation expenditures include: limiting the use of CFCs, adopting lower emitting fuels, fostering reforestation, raising the height of bridges, and moving water and waste treatment facilities.

A full accounting of benefits and cost should not be limited to a simple summing of individual adverse impacts or direct mitigation expenditures. Rather, a thorough analysis should reflect a general equilibrium framework with full social accounting. Moreover, the ancillary benefits or effects of any emissions mitigation strategy must also be incorporated in the benefit/cost calculation. These include the health and/or welfare benefits (due to a reduction in criteria air pollutants) or the energy security benefits of an energy saving measure and, conversely, the potentially lower agricultural yields as a result of the reduced CO₂ fertilization.

II. Benefits Estimates

Nordhaus derives many of his calculations of the economic impacts of climate change from a recent EPA report (1989) that developed estimates of the physical consequences associated with an assumed doubling of CO₂-equivalent concentrations of GHGs in the atmosphere. He concludes that, "[f]or the bulk of economic activity, non-climate variables like labor skills, access to markets, or technology swamp climatic considerations in determining economic efficiency" (1990, p. 17). By his estimation, the economic sectors likely to be

most severely affected by climate change are farming and fisheries. Nearly 90 percent of the economy is characterized as being negligibly affected by climate change. Nordhaus estimates that damage due to a 3°C warming in the middle of the next century is likely to equal one-quarter percent (0.25 percent) of annual U.S. national income. He allows that "inadequately studied or inherently unquantifiable" effects may raise the total to 1 percent of income, with an upper bound "unlikely to be larger than 2 percent of total output" (p. 20).

A review of Nordhaus' estimates suggests that the damages of global climate change may be larger in dollar amount in the categories he has monetized. Further, the damages may be broader in their scope, potentially affecting numerous other categories that he has neither provided dollar estimates for nor considered in a qualitative manner. Specifically, economic impacts due to sea-level rise have been revised to include an empirical study of the value of lost land, to add the value of land to the cost of constructing protection for sheltered areas, and to correct a computational error in the EPA report used in the cost calculation of protecting open coasts (see J. G. Titus et al., 1991). The combined effect of these revisions is an approximate doubling of Nordhaus' estimate of the annual economic effects of sea-level rise to \$10.6 billion (which also has the effect of doubling GNP losses from 0.25 to 0.5 percent). Both agricultural impacts and the effects of climate change on electricity operating and capital cost could also be substantially higher than Nordhaus allows (using high-end estimates from the EPA report could raise impacts in Nordhaus' categories to as much as 1.5 percent GNP).

Beyond reestimating the damage in categories that Nordhaus allows may be severely affected, a larger concern is for omitted or poorly defined categories. Some are relatively easy to measure and inherently monetizable, others are difficult to quantify and can best be weighed using some mix of quantitative and qualitative analysis. Still other categories such as preservation of coastal wetlands, old growth forests and biological diversity, with their attendant option

and existence values, are difficult if not impossible to quantify. All of which argues for more research on benefits and a plea to avoid the simple conclusion that if we cannot measure the benefits with precision, they effectively have zero or low value to society.

III. Comprehensive Approach

While much attention has focused on CO₂ limits in the context of climate change mitigation strategies, it is but one (albeit an important one) of a number of GHGs. Others include methane (CH₄), chlorofluorocarbons (CFCs), nitrous oxides (N₂O), and ozone (O₃). Of additional concern are gases that are not themselves GHGs, but are chemically active in the atmosphere and influence the concentration of GHGs. Examples of these are volatile organic compounds (VOCs), carbon monoxide (CO), and nitrogen oxides (NO_x). In addition, HCFCs which are important interim CFC substitutes should be included in a thorough analysis.

A comprehensive approach to GHG mitigation policy, at a minimum, should 1) encompass all direct and indirect GHGs, and 2) account for differences in the relative contribution of each gas to global warming. As to the second point, the Intergovernmental Panel on Climate Change (IPCC) (1990) has recently published "Global Warming Potential" (GWP) values for numerous GHGs. The GWP values allow emissions of GHGs to be expressed in a common metric ("carbon equivalents") that incorporates the differences in radiative forcing effect among GHGs and integrates that effect over the gases' atmospheric lifetimes, accounting for varying decay profiles and interactive (indirect) effects.

The IPCC GWP values relate trace gas emissions to a single physical effect—radiative forcing. A more elaborate index could be expanded to reflect the comparative effect of trace gas emissions on all physical phenomena and then translate those into socioeconomic impacts. Such an index has been suggested as a "Global Change Index." (See John Reilly, 1990; Richard Stewart and Jonathan Wiener, 1990.) One must ac-

TABLE 1—U.S. GREENHOUSE GASES EMISSIONS PROJECTIONS WITH CURRENT POLICY COMMITMENTS^a

GHG	1987	2000	2010
CO ₂	1310	1453–1503	1498–1627
CH ₄	235	208	212
VOCs	72	48	50
NO _x	218	199	235
CO	52	45	46
N ₂ O	74	74	74
CFC	367	256	188
Total	2328	2283–2332	2303–2430

Source: Cristofaro; Cristofaro and Scheraga; and Jorgenson and Wilcoxon.

^aCarbon equivalents, millions of tonnes; 100-year GWP factors.

knowledge, however, that a comprehensive approach to climate change mitigation (using the GWPs or a more elaborate index) faces numerous practical problems, including difficulties in inventorying gases and accounting for the locational, temporal, and source-specific nature of many of the direct and indirect GHG effects.

IV. Emissions and Cost Estimates

Table 1 presents estimates of total U.S. GHG emissions given current governmental policies for the years 1987, 2000, and 2010. The difference between the low and high estimates is due strictly to different CO₂ projections: the high CO₂ estimate is derived from the official U.S. submission to the IPCC, and the low CO₂ estimate is derived from the recent work of Dale Jorgenson and Peter Wilcoxon (1990). Subject to a long list of caveats discussed in the referenced papers, (weighted) U.S. emissions of all GHGs are projected to be approximately the same in the year 2000 as in 1987 (see Alexander Cristofaro, forthcoming; Cristofaro and Joel Scheraga, 1991). For the year 2010, estimates range from a slight increase (4 percent) to virtually no change. The base case scenarios indicate that an increase in CO₂ emissions (14–24 percent) will be offset by decreases in other gases covered by the Montreal Protocol (primarily CFCs) and the Clean Air Act.

Given these base case emissions scenarios, the question for policymakers is: What can be learned about the costs of meeting various GHG goals?

If the mitigation goal is to cap the overall weighted level of GHGs during the next 20 years (including so-called Montreal Protocol gases), only a limited amount of additional government intervention may be needed because current government energy and environmental policies are sufficient to keep total GHG emissions (expressed as carbon equivalents) below 1987 levels in the year 2000. In 2010, total emissions are projected to be only 4 percent above current levels in the high scenario, and essentially the same as 1987 levels in the low scenario.

If, instead, the goal is to stabilize or reduce CO₂ emissions or to reduce non-Montreal Protocol GHGs, additional control programs would be necessary. The rate of increase of CO₂ in the low scenario, however, is less than the rate of projected U.S. population growth (currently 0.7 percent per year). The rate of increase of CO₂ in the high scenario is only slightly more than projected population growth. Thus, existing government programs may be sufficient or nearly sufficient to limit per capita CO₂ emissions to today's levels in the years 2000 and 2010. More ambitious goals would require larger reductions.

A number of additional control opportunities exist that have not been systematically analyzed in the context of current policy. Methane is 21 times more potent than CO₂ in its radiative forcing contribution over a 100-year period and, thus, methane control may be an important component of GHG policy. Control opportunities include improved management, recovery, and fuel use of methane from animal waste, coal mines, solid waste landfills, and the anaerobic decomposition of sewage sludge. NO_x is 40 times more potent than CO₂. Although the 1990 Clean Air Act will result in over 2 million tons of NO_x reductions, other technically feasible NO_x control opportunities exist. CO is 8 times more potent than CO₂. Tighter tailpipe standards and oxygenated fuel requirements mandated by the 1990 Clean Air Act could be broadened or other-

wise strengthened. For all methane, NO_x, and CO options, additional analysis of total program cost and benefits, including complete fuel cycle analysis and health and welfare benefits, should be conducted before controls are implemented.

A number of studies have looked at the costs of controlling CO₂. The estimates of control costs vary widely, largely due to different assumptions regarding emission baseline levels, reduction targets, GDP and population growth rates, rates of technological change, estimates of capital mobility and malleability, ease of factor substitution, relative energy prices and, importantly, underlying model structure. Technology-driven end-use models generally project significant potential for cost-effective energy savings (i.e., net savings). A U.S. Department of Energy (1990) white paper projects the potential for cost-effective efficiency gains (reductions in demand) of 14 percent below otherwise projected consumption levels in 2010. Other "bottom-up" studies project larger cost-effective or "technological potential" savings. Bottom-up approaches lead to interesting behavioral, policy, and analytic questions: How much of the cost-effective potential can be captured? How best to capture it?

The CO₂ control cost estimates exist from a variety of macroeconomic studies. Generally, these studies, which include short- and long-run models and differ widely in their methodologies and input assumptions, estimate positive control costs. Typical results estimate that a 10 to 20 percent reduction from base levels will require a carbon tax of between \$10 and \$75 per ton, and GNP will be 0.2 to 4.0 percent lower than otherwise. Among the studies surveyed by Nordhaus is the study by Jorgenson and Wilcoxon (J/W). The J/W study is based on a 35-sector econometric model of the U.S. economy estimated through 1985 with endogenous technical change. Economic growth is an output rather than input of this model, and is projected to grow at a 2 percent annual rate over the 1990–2000 period, with a slight decrease thereafter. Preliminary results from J/W of a carbon tax analysis indicate that emissions can be maintained at 1990

levels through the imposition of a phased carbon tax that rises to a level of approximately \$17 per ton in 2020. (Assumptions would have to be altered to develop credible projections beyond 2020. The \$17 per ton estimate, in part, reflects an assumed high degree of capital malleability in the utility sector; the model is being revised to address this aspect, and adjustments will likely result in cost estimates somewhat higher.)

A carbon tax, absent accommodative measures, and depending upon its timing and scope, could have inflationary and/or recessionary impacts. The use of a "tax shift" mechanism could help minimize the net social cost of CO₂ control; revenues generated through a carbon tax could be rebated or offset by reductions in income and/or business taxes. Importantly, selection of the precise measure used to recycle the carbon tax can have a significant differential effect. To illustrate this phenomenon, consider a recent study by Data Resources, Inc. (DRI/McGraw Hill, 1991) that analyzed the differential macroeconomic impacts of a gasoline tax sufficient to maintain light duty vehicle carbon emissions at 1989 levels through 2010. In one scenario analyzed, the gasoline tax revenues were accompanied by a reduction in personal income taxes, resulting in a GNP loss of 0.4 percent in the year 2000. In another scenario, the gasoline tax revenues were accompanied by a reduction in payroll taxes (paid by businesses), resulting in a GNP virtually unchanged from baseline levels.

The above two scenarios highlight the possibility of addressing an environmental externality through a Pigouvian tax while minimizing the adverse macroeconomic effects (caused by the resulting change in fiscal policy) through the adoption of other appropriate policies. A GHG mitigation policy that shifts the tax base toward the incorporation of an environmental externality and away from other business taxes (in some sense, trading a good tax for a bad) is likely to minimize the impact on total national income, although tax burdens and economic impacts on different sectors of the economy will vary.

V. Conclusion

Apart from the scientific uncertainties, a complex set of issues surround the economic analysis of climate change. The calculation of net benefits related to optimal policy strategies hinges on at least four key factors: 1) the breadth of quantifiable damages considered; 2) the number of GHGs for which practical control options are devised (and ancillary net benefits included); 3) the consideration of base case growth, technological damage, capital malleability, and related factors; and 4) the extent to which other policies (i.e., tax recycling/shifting) can be used to offset the macroeconomic effects of GHG policies. A great deal more research is needed—especially an attempt to look at both costs and benefits in an international context. I encourage other economists to engage the issue.

REFERENCES

- Cristofaro, Alexander, "The Cost of Reducing Greenhouse Gas Emissions in the United States," *Proceedings of Global Climate Change Conference*, Washington: Center for Environmental Information, forthcoming.
- _____ and Scheraga, Joel D., "Policy Implications of A Comprehensive Greenhouse Gas Budget," *Forum For Applied Research and Public Policy*, forthcoming, Fall 1991.
- Jorgenson, Dale W. and Wilcoxon, Peter J., "The Cost of Controlling U.S. Carbon Dioxide Emissions," preliminary draft, September 1990.
- Nordhaus, William D., "To Slow or Not To Slow: The Economics of the Greenhouse Effect," Cowles Foundation discussion paper, 1990.
- _____, "The Cost of Slowing Climate Change: A Survey," *The Energy Journal*, forthcoming 1991.
- Reilly, John, "Climate Change Damage and the Trace Gas Index Issue," draft paper, November 1, 1990.
- Stewart, Richard B. and Wiener, Jonathan B., "A Comprehensive Approach to Climate Change," *American Enterprise*, Nov.-Dec.

- 1990.
- Titus, James G. et al., "Greenhouse Effect and Sea Level Rise: Potential Loss of Land and the Cost of Holding Back the Sea," *Coastal Management*, forthcoming 1991.
- DRI/McGraw-Hill, "An Analysis of Public Policy Measures To Reduce Carbon Dioxide Emissions From The U.S. Transportation Sector," January 1991.
- Intergovernmental Panel on Climate Change, *Climate Change: The IPCC Scientific Assessment*, Cambridge: Cambridge University Press, 1990.
- U.S. Department of Energy, *Energy Efficiency: How Far Can We Go*, prepared by Oak Ridge National Laboratory, Oak Ridge, TN, 1990.
- U.S. Environmental Protection Agency, *The Potential Effects of Global Climate Change on the United States*, Washington: USGPO, 1989.

A Sketch of the Economics of the Greenhouse Effect

By WILLIAM D. NORDHAUS*

Over the last decade, scientists have studied extensively the greenhouse effect, which holds that the accumulation of carbon dioxide (CO₂) and other greenhouse gases (GHGs) is expected to produce global warming and other significant climatic changes over the next century (see Stephen Schneider, 1989). The present study presents an economic approach for analyzing policies to slow climate change.

I. Scientific Background

In weighing climate change policies, the prospects for global warming and the linkage between human activities and the emissions of GHGs form a key building block. The major GHGs are carbon dioxide, methane, nitrous oxides, and chlorofluorocarbons (CFCs). In this analysis, I translate each GHG into its "CO₂ equivalent" in total warming potential. Current emissions are about 8 billion tons of CO₂ equivalent per year (carbon weight). CO₂ is the dominant GHG, currently contributing about 80 percent of the total potential global warming of the major GHGs.

On the basis of climate models and climate history, scientists expect that a doubling of CO₂ (or its radiative equivalent) will in equilibrium lead to an increase in the global mean surface temperature of 1 to 5 degrees Centigrade (C), an increase in precipitation and evaporation, a small rise in sea level, and a number of other more uncertain effects.

A complication in studying the climatic impact of rising GHGs involves the time delay of several decades, arising because of

the thermal inertia of the oceans, in the reaction of climate to increasing atmospheric concentrations. In the model used here, I simplify by assuming that the temperature-adjustment process takes the following form:

$$(1) \quad \dot{T}(t) = \alpha\{\mu M(t) - T(t)\}$$

$$(2) \quad \dot{M}(t) = \beta E(t) - \delta M(t)$$

where dots over variable represent time derivatives and t = time; $T(t)$ = increase in global mean surface temperature since mid-nineteenth century from accumulation of GHGs (°C); $M(t)$ = anthropogenic atmospheric concentration of CO₂ equivalent GHGs (billions of tons of CO₂ equivalent); $E(t)$ = anthropogenic emissions of CO₂-equivalent GHGs (same units as M); μ = .005 = linearized equilibrium increase in global mean temperature in response to increasing CO₂ equivalent concentration (°C per billions tons carbon); α = .02 = delay parameter of temperature in response to radiative increase (per year); β = .50 = airborne fraction of CO₂ equivalent emissions; and δ = .005 = annual rate of removal of CO₂ equivalent from the atmosphere (per year).

Equation (1) is a simple first-order equation often used to represent climate dynamics, while equation (2) is a simplified two-box diffusion model of the carbon cycle. These equations are calibrated or estimated from current models or historical data.

II. Economic Aspects of Greenhouse Warming

Greenhouse warming is the granddaddy of all public goods, involving climate change over the indefinite future. Because of the climate externality, the production of greenhouse gases will differ from the efficient level. We can analyze the costs and benefits of the greenhouse effect and policies in terms of two fundamental functions: the

*Yale University and the Cowles Foundation, New Haven, CT 06520. This paper is an abridged version of my longer study (1990b), and I am grateful for insightful comments from many people, with particular thanks to Alan Manne, Tom Schelling, James Sweeney, and Paul Waggoner. This research was supported in part by the National Science Foundation.

greenhouse damage function that describes the costs to society of the changing climate, and the *abatement cost function* that describes the costs that the economy pays to slow greenhouse warming. I analyze here efficient strategies, those that maximize overall net economic welfare ("Green National Product").

Economic theory tells us that, in a competitive economy with no other externalities and where controls are efficiently designed, certain properties hold. To begin with, the first units of GHG reduction are virtually free because of the zero market price on the GHG emissions. Second, we know that the cost function increases in the level of abatement.

Next examine the greenhouse damage function that measures the cost to the economy of higher levels of GHGs (measured relative to some baseline). In contrast to the cost function, we know little about the shape of the damage function. We suspect that greenhouse warming will hurt the global economy, but because of the fertilization effect of CO₂ or the attractiveness of warm climates, the greenhouse effect might on balance actually be economically advantageous. In addition, the costs of climate changes are likely to be sensitive to the speed of climate change.

I next present a stylized model of the relationship between economic growth and climate change which incorporates the dynamics of climate change and of investing to slow climate change. I study the impact of greenhouse policies upon an idealized global economy in the middle of the next century. The key assumption is that the economy is in "resource steady state," which signifies that all *physical* flows in the global economy are constant although the real *value* of economic activity may be increasing. I allow for "balanced resource-augmenting technological change" at rate h .

In the steady state, per capita consumption is given by

$$(3) \quad c(t) = y^* e^{ht} \{ g[E^*] - \phi[T^*] \}$$

In this equation, the asterisks indicate steady-state values. The new variables are $c(t)$ = per capita consumption at time t ; y^*

is a constant; $y(t) = y^* e^{ht}$ = per capita output before any emissions reduction and with no climate damage; $g(E)$ = cost of emissions reduction; and $\phi(T)$ = damage function from greenhouse warming.

It is assumed that it is desirable to maximize a social welfare function that is the discounted sum of the utilities of per capita consumption. This is in essence a standard optimal growth model extended to include climatic variables. In the model used here, the critical parameter is $r - h$, the difference between the discount rate on goods (r) and the growth rate of the economy. This will be relevant because, while we discount future damages at r , in our resource steady state the damages will be growing at the rate of economic growth (h).

To calculate the optimal level of emissions reduction in our simple model, let us perform a variational experiment of increasing emissions by ΔE in period 0. A fair amount of algebra will show that the optimal level of emissions is given by

$$(4) \quad g'(E^*) = \mu \beta \phi'(T^*) \Gamma$$

where Γ = a present-value factor = $\alpha / [(r + \delta - h)(r + \alpha - h)]$. Equation (4) states that the optimal degree of reduction of GHGs comes where the marginal current cost of reducing GHG emissions equals the present value of the marginal damage from higher concentrations. The present-value factor Γ can be interpreted as the number of years, in present value, of equilibrium-CO₂-doubling climate damage that occur when a one-shot concentration increase, equal to the initial CO₂ concentration, takes place at time zero. In the section that follows, I estimate the parameters of this equation and then derive the optimal rate of emissions reduction for GHGs. (A preliminary analysis of a non-steady-state trajectory incorporating several regions and growth of emissions is contained in my 1990a paper.)

III. Empirical Estimate of the Optimal Strategy

I have elsewhere estimated the economic parameters underlying the optimal condition in equation (4). These estimates are

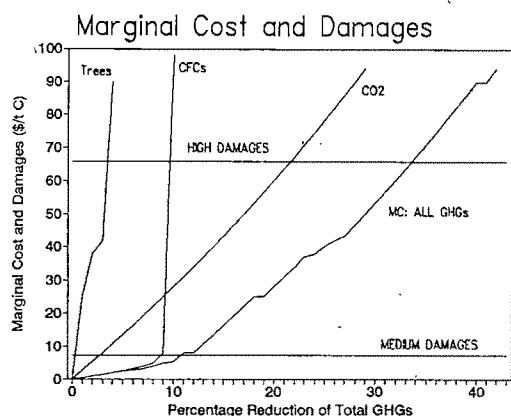


FIGURE 1

highly uncertain, and the reservations are given in the background documents.

There are numerous estimates, particularly for CO_2 , of the cost of reducing GHGs (see the survey in my forthcoming article). Figure 1 shows the overall marginal cost curve and the cost curves for different GHGs derived in that study. With current annual emissions of 8 billion tons of CO_2 equivalent, these data suggest that a modest reduction of GHG emissions can be obtained at low cost. After a 10 percent reduction, however, the curve rises as more costly measures are required. A 50 percent reduction in GHG emissions is estimated to cost almost \$200 billion per year in today's global economy, or around 1 percent of world output. This estimate is understated to the extent that policies are inefficient or are implemented in a crash program.

Studies of the impacts of climate change are sparse. They cover few countries and have major methodological shortcomings, so our estimates of the impact of greenhouse warming are only suggestive. Studies indicate that of the major industries, agriculture is the one most likely to be severely affected by climate change, although the sign of the impact is uncertain. Sea-level rise will be costly to coastal regions, but the impact is likely to be very gradual. Other affected sectors are construction, energy, recreation, and water systems, but the impacts here are

likely to be imperceptible compared to the background noise for many decades.

For the numerical estimates used here, I examined the impact of climate change coming from an equilibrium doubling of CO_2 equivalent atmospheric concentrations. This is assumed to produce a 3°C rise in global mean surface temperature along with the associated changes in climate. In an uncontrolled economy, this is likely to arrive in the second half of the next century. In the aggregate and in terms of those variables we have been able to quantify, I estimate that the net economic damage from a 3°C warming is likely to be around $\frac{1}{4}$ percent of national income for the 1981 sectoral distribution of the United States (we call this the "low" estimate). This figure is clearly incomplete, for it neglects a number of areas that are either inadequately studied (nonmarket sectors) or difficult to quantify (value of species loss or aesthetic values). We might raise the number to around 1 percent of total global income to allow for these unmeasured and unquantifiable factors, although such an adjustment is purely *ad hoc* (this is the "medium" estimate). It is not possible to give precise error bounds around this figure, but my judgement is that the overall impact upon human activity is unlikely to be larger than 2 percent of total output (this is the "high" estimate).

IV. An Efficient Policy for Slowing Greenhouse Warming

We can now provide estimates of an efficient policy for slowing greenhouse warming, where this is described in equation (4). In this analysis, it is assumed that the baseline policy is one in which there are no greenhouse policies in place.

Let us begin tabulating in Table 1 the calculated costs and damages that are drawn from the findings above. Column (1) shows the percentage reduction in GHGs from an uncontrolled level. Columns (2) and (3) shows the costs of GHG reductions from Figure 1. The final column displays the estimated total discounted damages associated with the given level of reduction of GHG emissions, this figure being based on the middle impact estimate.

TABLE 1—CALCULATION OF COSTS AND BENEFITS FOR DIFFERENT LEVELS OF REDUCTION OF GREENHOUSE GAS EMISSIONS

(1)	(2)	(3)	(4)
0	.0	.0	.0
1	.5	.04	.59
2	1.0	.12	1.2
3	1.5	.24	1.8
4	2.0	.40	2.3
5	2.6	.61	2.9
10	5.3	2.2	5.9
11 ^a	8.0	2.9	6.4
15	16.3	6.8	8.8
25	40.2	30.7	14.7
50	120.0	191.0	29.3
75	285.0	581.0	44.0

Source: For both costs and benefits, calculations use 1989 levels of world greenhouse gas emissions, prices, and world output. Cost estimates shown in Figure 1. Background data are presented in my 1990b paper.

Note: Col. 1 = Reduction of GHG emissions (as percent of base level); Col. 2 = Marginal cost of reduction (\$ per ton CO₂); Col. 3 = Total cost of reduction (bill. of \$/yr); Col. 4 = Total benefit of reduction (bill. of \$/yr).

^aMost efficient level of control of GHG emissions for medium damage level.

The efficient level of GHG reduction is shown in Table 1 for the middle level of damages and for a discount rate that is 1 percent above the growth rate (i.e., $r - h = .01$ per year). This estimate corresponds to the middle damage estimate of \$7.30 per ton of CO₂ equivalent. Equating the marginal damage with the marginal cost leads to an efficient level of control of 11 percent of GHG emissions. At the efficient control level, the total global annual cost of reducing emissions is around \$3 billion while the total global annualized benefits are estimated to be around \$6 billion.

The same outcome is illustrated in Figure 1, which puts together the empirical marginal costs and damage curves. The horizontal axis shows the percentage reduction in GHGs. The curve marked "MC: All GHGs" is my estimate of the marginal cost of GHG reduction. The horizontal curves marked Medium and High Damage correspond to marginal damage estimates of \$7.30

and \$66 per ton of CO₂ equivalent; these correspond to annual damages of 1 and 2 percent of total output, respectively. The first figure uses the middle discount rate of $(r - h)$ equal to 1 percent, while the second uses a value of $r - h$ of 0.

The efficient policy is found at the intersection of the relevant damage curve with the marginal cost curve. The medium case leads to a reduction of 11 percent of GHG emissions. At the low damage estimate, not shown, there is very little GHG emission reduction for its own sake. At the extreme, high damage estimate, about one-third of total GHG emissions would be efficiently reduced.

Figure 1 also allows us to determine the contribution of different GHGs to the total reduction. For the medium damage estimate, virtually none comes from trees, 2 percent from the reduction of CO₂ and 9 percent comes from the reduction in CFC emissions. All options suggest a significant reduction in the use of CFCs and that little can be realized through forestry options. The main difference among the policies is the extent to which CO₂ emissions are reduced.

V. Conclusions

This study examines an economic approach to policies concerning greenhouse warming. It uses an approach in which the optimal degree of control of GHGs is set at that point where the marginal costs of control are equal to the discounted marginal damages from climate change. For the low damage function (that includes only identified costs and uses a middle discount rate), I estimate the marginal damage of greenhouse gases to be about \$1.80 per ton of C in CO₂ equivalent, which implies very little CO₂ abatement. For the medium damage function, that assumes damage from greenhouse warming of 1 percent of GNP, the cost is reckoned at \$7.30 per ton carbon; in this case, the efficient reduction is 11 percent of total GHG emissions. In the high damage case, with damages taken to be 2 percent of total output and with no discounting, GHG emissions are reduced by about one-third.

Second, the appropriate level of control depends critically upon three central parameters of the climate-economic system: the cost of control of GHGs, the damage to the human societies from greenhouse warming, and the time dynamics as reflected in the rate of discount of future goods and services along with the time lags in the reaction of the climate to emissions. The efficient degree of control of GHGs would be essentially zero in the case of high costs, low damages, and high discounting; by contrast, in the case of no discounting and high damages, the efficient degree of control is close to one-third of GHG emissions.

Finally, it should be emphasized that this analysis has a number important oversimplifications. It greatly simplifies many of the intricate economic and climatic complexities by taking a global view of economic activity and a simple dynamic specification of emissions, concentrations, and economic growth. Moreover, this study bases the economic damage assumptions upon the 1981 sectoral composition of the U.S. economy and assumes that this composition will hold for the global economy in the mid-twenty-first century. In addition, it ignores other routes for investing society's resources (such as factories, education, research, and health) and focuses on a single tradeoff between future and present consumption. Furthermore, the calculations omit other potential market

failures, such as ozone depletion or air pollution, that might reinforce or weaken the logic behind greenhouse-gas reduction. And finally, it ignores the issues of uncertainty, in which risk aversion and the possibility of learning may modify the stringency and timing of control strategies. Notwithstanding these simplifications, the approach laid out here may help clarify the questions and identify the scientific, economic, and policy issues that must underpin any rational decision.

REFERENCES

- Nordhaus, William D., (1990a) "An Intertemporal General-Equilibrium Model of Economic Growth and Climate Change," paper presented to a Workshop on Economic/Energy/Environmental Modeling for Climate Policy Analysis, Washington, D.C., October 1990.
- _____, (1990b) "To Slow or Not To Slow: The Economics of the Greenhouse Effect," Cowles Foundation discussion paper, 1990.
- _____, "The Cost of Slowing Climate Change: A Survey," *The Energy Journal*, forthcoming 1991.
- Schneider, Stephen H., "The Greenhouse Effect: Science and Policy," *Science*, February 10, 1989, 243, 771-81.

GENDER AND PRODUCTIVITY

The Role of Off-the-Job vs. On-the-Job Training for the Mobility of Women Workers

By LISA M. LYNCH*

The transition from school to work is typically a period in which many young workers experience a wide range of different jobs and experience some of their most rapid wage growth over their working life. Robert Hall (1982) has estimated that the first 10 years of an individual's working career will include approximately two-thirds of all lifetime job changes. Robert Topel and Michael Ward (1988) found that over half of young male new entrants held six or more jobs over the first 10 years of their work experience. Only one young male worker in twenty held a single job for 10 years in their sample. All of this suggests that young workers' early years in the labor market involve several employment transitions.

The purpose of this paper is to examine for young workers in their first years of work the determinants of leaving an employer. In particular, this paper focuses on the role of different types of training on the probability of leaving an employer. In previous work (1990) I have examined the impact of private-sector training on the determination of wages and wage growth of young workers and reached the following conclusions. First, formal company-provided on-the-job training (ON-JT) appears to be highly firm-specific in the United States and, therefore, is not portable from employer to employer. Company-provided training raises

wages in the current job but has no effect on the wages earned in subsequent employment. Second, formal off-the-job training (OFF-JT) received from "for-profit" proprietary institutions has little effect on the wages earned on the current job, but it does raise the expected wage in subsequent employment. Finally, there are important differences by race, gender, and education level in the probability of receiving different types of formal training and in the impact this training has on wages and wage growth.

These findings have several implications for the impact of training on mobility. One implication is that if company-provided training is primarily firm-specific, then the probability of leaving an employer should decline if a young worker has experienced some ON-JT. An additional implication is that if a worker participates in an OFF-JT program, it appears that the worker should be more likely to leave the current employer. In this case, OFF-JT allows a young worker to change career paths and find a "better match." Using data from the *National Longitudinal Survey Youth (NLSY)* cohort, this paper examines in detail the factors that influence the probability of new entrants leaving their first job, including the differential effects of company-provided training, apprenticeships, and training from for-profit proprietary institutions.

I. The Theoretical Framework

There are a variety of explanations of why young workers change their employment status so often in the early years of their careers and then seem to settle down into more stable employment. In the unionized sector, where seniority rules determine lay-

*MIT Sloan School of Management, Cambridge, MA 02139, and NBER. Thorough research assistance was provided by Pamela Loprest and funding for this research was provided by the Department of Labor, grant number E-9-J-9-0049. The views expressed here do not necessarily represent the official position of the Department of Labor.

off policies, young workers are more at risk of being laid off in a downturn. Even in the nonunionized sector, many firms use seniority as a major determinant of whom to lay off in a period of falling demand.

There are other explanations of the higher turnover rates of young workers, however, that have little to do with the state of demand. The three main theoretical explanations include job search, job matching, and on-the-job training. Job search theory, as detailed by Steven Lippman and John McCall (1976), states that information about where to find a job and the nature of that job are difficult to acquire, especially for younger workers. Workers will accept employment and remain in that job as long as the wage paid in that job exceeds the alternative wage. Therefore, workers who earn more relative to their alternative wage are less likely to quit.

An alternative explanation of turnover behavior can be found in Boyan Jovanovic (1979a,b; 1984). In the Jovanovic learning model, both workers and firms "learn" about the unobserved characteristics of each other over time. As tenure increases, the quality of the job match is revealed as firms observe workers' actual productivity and workers discover the nonpecuniary aspects of their job. In this model there are two countervailing forces for the relationship between tenure and the probability of leaving an employer. On the one hand, "better" workers remain with employers longer, leading to negative duration dependence in the probability of leaving a job. On the other hand, as "bad" matches are revealed, the turnover probability will rise over time.

The process of ON-JT within the human capital model as described by Jacob Mincer (1974) implies that as workers acquire firm-specific training, their productivity and, consequently, wages will rise. Therefore, the probability of leaving an employer will fall with training and tenure since the wage will rise relative to the alternative wage. In addition, employers will be less likely to lay off those workers in whom they have invested in specific skills. However, if most of the initial training for young workers is general, there will be either no effect on the quit

probability or the quit probability may even rise.

All of these theories are not mutually exclusive and clearly some combination of all of these factors influences the probability of a young worker remaining with an employer. Consequently, it is not the purpose of this paper to distinguish between different theories. Rather, it would be more useful if precise data on employment spells and training could be found, to try to establish the links between different types of training and turnover behavior.

There have been relatively few empirical studies that have attempted to examine the role of training, demand, and other factors in predicting the probability of leaving an employer. This is primarily due to the lack of accurate data on the timing of private sector training and the lack of detailed employment histories for workers. Recent exceptions include R. Mark Gritz (1988) and Mincer (1988). Gritz uses data from the early years of the *NLSY* and finds that private-sector training (not distinguishing between different sources of training) increases the amount of time in total employment for females, but decreases the amount of time males were employed. Unfortunately, most of the training data he uses are taken from the period before the employment history begins. Mincer uses data on training and mobility from the *Panel Study of Income Dynamics*. The training variable comes from the answer to the following question in the 1976 and 1978 interviews: "On a job like yours how long does it take the average new person to become fully trained and qualified?" While this is potentially a very broad measure of training, it does not measure how much training has actually occurred for the specific respondent. In addition, it captures training information for the current job, not previous employment.

Using data from the *NLSY*, it is possible to examine in more detail than has been possible in the past, the role of training, the general state of demand, and other personal characteristics in determining turnover. The probability of leaving employment (for whatever reason) is also known as the haz-

ard rate or failure rate in renewal theory. The hazard rate or turnover probability can be expressed as

$$(1) \quad h(t) = g(t) dt / (1 - G(t)),$$

where $g(t)dt$ is the probability of leaving an employer between time t and $t + dt$, $1 - G(t)$ is the probability of being employed at time t , and t is the duration of the current spell of employment. In this paper, the following Cox proportional hazards model is used:

$$(2) \quad h(t; z) = h_0(t) e^{zB},$$

where $h_0(t)$ is an arbitrary and unspecified baseline hazard function and z is a vector of characteristics including training. The Cox model is convenient for dealing with right censoring and it is nonparametric in the sense that it involves an unspecified baseline hazard instead of making further distributional assumptions such as those required for the Weibull or log-logistic hazard. However, this means that it will not be possible to measure whether or not there is negative or positive duration dependence in employment, but this is not a key focus of this paper.

In a model of the role of training in the probability of leaving an employer, it is important to be able to allow training to occur over time with the employer. Allowing for covariates such as training to be time dependent implies

$$(3) \quad h(t; z(t)) = h_0(t) e^{z(t)B},$$

where $z(t)$ is a vector of all fixed and time varying covariates.

As discussed in D. R. Cox and D. Oakes (1984), the components of the vector $z(t)$ can be divided into the following three categories of variables: treatments that vary with time; intrinsic properties of individuals/jobs that are time invariant; and exogenous time-varying variables.

Obviously the different types of private-sector training are the treatment variables of interest. Examples of time-invariant per-

sonal and job characteristics include gender, race, education, occupation, industry, union status, location of the job in an urban area, and whether or not the respondent is disabled. Time-varying exogenous variables include the local unemployment rate, marital status, and the number of children.

II. The Data

The *NLSY* is a survey of 12,686 males and females who were 14 to 21 years of age at the end of 1978. These respondents have been interviewed every year since 1979 on all aspects of their labor market experience. It is possible to construct a detailed weekly employment history for all the respondents since January 1, 1978. This can be matched with a similar event history of schooling and training from the private sector. The data on types of training received (other than governmental training or schooling) are some of the most comprehensive data available on private-sector training. Respondents were asked about what types of training they had received over the survey year (up to 3 spells) and the dates of training periods by source. Potential sources of training include business college, nurses programs, apprenticeships, vocational and technical institutes, barber or beauty schools, correspondence courses, and company-provided training.

The training data are divided into the variables: company training (*ON-JT*); apprenticeships (*APT*); and training obtained from for-profit proprietary institutions outside the firm (*OFF-JT*). The variable *OFF-JT* includes training obtained from business courses, barber or beauty school, nurses programs, vocational and technical institutes, and correspondence courses. All of these types of training programs are independent from training received in a formal regular schooling program that is included in the schooling variables. However, the questions ask about only those spells of training that lasted at least 4 weeks (they did not have to be full time). This suggests that the *NLSY* measure of training is more likely to capture formal training spells than informal on-the-job training.

For this analysis, a subsample of the 12,686 respondents has been selected. I have excluded the 1280 respondents in the military subsample from the analysis. I have also deleted any respondent who had completed school before the 1979 interview year. The final sample is a pooled sample of young workers who have left school and not returned to school for at least 4 years (permanently out of school). Therefore, this sample is made up of 5 waves of school leavers: those who left in 1979, 1980, 1981, 1982, and 1983. In addition, the respondents had to have obtained a job in the first year after permanently exiting school. The estimated hazard models the determinants of the turnover probability for the first job after leaving school permanently for this sample.

Almost three-quarters of the sample leave their first employer during the first 4 years after school. The average duration of employment (including those still employed after 4 years) is about a year and a half. Almost 17 percent of the sample experienced some form of formal training in their first job, but the distribution of this job training by source varied substantially by demographic group. College graduates were much more likely to have received some form of ON-JT while those with just a high school diploma were more likely to have participated in some form of OFF-JT. Women were more likely than men to have received some form of OFF-JT, but there was little difference in the probability of receiving ON-JT by gender (not controlling for other factors).

III. The Results

The results obtained from estimating the Cox proportional hazard with time-varying covariates are presented in Table 1. The time-invariant intrinsic characteristics of the individuals/jobs in Table 1, equation 1, that seemed to influence the probability of leaving an employer included being disabled, union status, race, and school level. Disabled respondents were more likely to leave their employer as were blacks and those with a high school degree or less. Being

TABLE 1—DETERMINANTS OF THE PROBABILITY OF LEAVING EMPLOYER

Variable	Eq. 1 All, N = 2522	Eq. 2 All, N = 2522	Eq. 3 Male, N = 1208	Eq. 4 Female, N = 1314
<i>Urban</i>	-.06 (-1.31)	-.04 (-0.95)	-.06 (-0.94)	-.03 (-0.37)
<i>No.</i>	.09 (1.59)	.09 (1.54)	-.08 (-0.71)	.20 (2.83)
<i>Children^a</i>	.22 (1.98)	.22 (1.96)	-.14 (-0.67)	.41 (3.01)
<i>Disabled</i>	-.22 (-3.44)	-.21 (-3.27)	-.36 (-2.84)	-.13 (-1.66)
<i>Married^a</i>	-.28 (-4.34)	-.27 (-4.16)	-.15 (-1.73)	-.44 (-4.40)
<i>Union</i>	.14 (2.41)	.11 (1.99)	.10 (1.30)	.10 (1.17)
<i>Black</i>	.05 (0.83)	.04 (0.60)	.03 (0.35)	.02 (0.26)
<i>Hispanic</i>	-.05 (-1.20)	-.06 (-1.32)	—	—
<i>Male</i>	.69 (8.51)	.67 (8.22)	.39 (3.48)	.93 (7.72)
<i>Less than H.S.</i>	.26 (4.16)	.23 (3.65)	.08 (0.81)	.31 (3.74)
<i>High School</i>	-.24 (-2.79)	-.24 (-2.73)	-.53 (-3.79)	-.04 (-0.32)
<i>College</i>	-.17 (-2.95)	-.18 (-3.13)	-.18 (-2.27)	-.16 (-2.02)
<i>Medium</i>	-.17 (-2.74)	-.17 (-2.83)	-.23 (-2.66)	-.10 (-1.17)
<i>Unrate^a</i>	-.40 (-2.62)	-.32 (-2.12)	-.27 (-1.19)	-.36 (-1.70)
<i>ON-JT^a</i>	.10 (1.51)	.11 (1.70)	.03 (0.27)	.61 (2.09)
<i>OFF-JT^a</i>	.03 (0.13)	.10 (0.48)	-.03 (-0.13)	.61 (1.21)
<i>APT^a</i>	—	-.64 (-10.04)	-.55 (-6.18)	-.79 (-8.44)
<i>Log Wage Diff.^a</i>	—	—	—	—
<i>Log Likelihood</i>	-14697.7	-14644.4	-6337.3	-6879.4

Notes: The *T* statistics are shown in parentheses. Equations also include dummy variables for year of entry.

^aDenotes time-varying covariates.

male or Hispanic had no differential effect on the expected duration of employment in the first job. Being employed in a job covered by a collective agreement or being a college graduate significantly lowered the probability of leaving the first employer. Of the time-varying exogenous covariates, the local unemployment rate was significant, implying that those who were in high unemployment areas were less likely to leave their employer. The hurdle for these youths

in high unemployment areas was most likely getting the job in the first place. In addition, those workers who were married were more likely to remain with their first employer. With regards to the training variables, those young people who had some formal *ON-JT* were much less likely to leave their employer while those who participated in some form of *OFF-JT* were more likely to leave. This seems to suggest that *ON-JT* is more firm-specific, while *OFF-JT* is more general. These findings are consistent with the results on training and wages reported in my earlier paper.

In equation 2 of Table 1, an additional variable is added that is the difference between the log of the current wage (that varies with time) and a log predicted wage. The predicted wage is obtained by using the estimated coefficients from a log wage equation for the starting wage for this sample. Those individuals who are being paid less than their predicted alternative wage are more likely to leave their employer as shown in equation 2. None of the previous findings are altered except that the degree of significance of *OFF-JT* improves.

In equations 3 and 4 the proportional hazard is reestimated for males and females. Now the results changed dramatically depending upon whether the worker is male or female. For example, it is striking to find that for men being disabled or just a high school graduate has no effect on the turnover probability. In addition, both *ON-JT* and *OFF-JT* are now insignificant. On the other hand, *ON-JT* increases the length of time in employment in the first job for women and *OFF-JT* increases their turnover probability. Having more children increases the probability that women will leave their employer, while marital status has no effect on the transition probability.

IV. Conclusions

While this paper has attempted to shed new light on the skill formation process of young workers and the consequences of this on their patterns of mobility, there are still many issues that remain unresolved. This paper has modeled the determinants of the

duration of the first job after school, not subsequent employment. As the *NLSY* age, future research should examine, for example, how some of the gender, race, and educational differences change over time. It would also be interesting to examine the hazard rates by broad industry and occupational categories. Finally, it would be important to see how robust the findings are after additional work is done to address the endogeneity issue for training.

Nevertheless, there is a story that emerges from the results in this paper for young workers and private-sector training. Company training in the United States is firm-specific, even for young workers in their first job. Young workers entering the labor market can receive both good and bad draws from the labor market. There are some workers who get a bad draw who appear to move to better employment by investing in off-the-job training. Those in good jobs are more likely to obtain on-the-job training that results in higher wages and a lower probability of leaving the firm. These effects are particularly strong for women.

REFERENCES

- Cox, D. R. and Oakes, D., *Analysis of Survival Data*, London: Chapman and Hall, 1984.
- Gritz, R. Mark, "The Impact of Training on the Frequency and Duration of Employment," mimeo., University of Washington, 1988.
- Hall, Robert, "The Importance of Lifetime Jobs in the U.S. Economy," *American Economic Review*, September 1982, 72, 716-24.
- Jovanovic, Boyan, (1979a) "Job Matching and the Theory of Turnover," *Journal of Political Economy*, October 1979, 87, 972-90.
- _____, (1979b) "Firm-Specific Capital and Turnover," *Journal of Political Economy*, December 1979, 87, 1246-60.
- _____, "Matching, Turnover, and Unemployment," *Journal of Political Economy*, February 1984, 92, 108-22.
- Lippman, Steven and McCall, John, "The Economics of Job Search: A Survey, Part I," *Economic Inquiry*, June 1976, 14, 155-89.
- Lynch, Lisa M., "Private Sector Training and

The Impact of Nonmarket Work on Market Wages

By JONI HERSCH*

It is frequently asserted that balancing a job and family responsibilities is more difficult for women than men. Support for this notion stems from evidence that women, in essence, work two jobs—one in the market and one at home. While the popular press has focused on the stress and frustration associated with the so-called “second shift,” work at home may also affect the labor market situation of women. The types and locations of jobs acceptable to women who assume heavy household responsibilities may be limited. In addition, nonmarket work may have a direct effect on earnings by reducing the amount of energy and effort available for market work.

Economists have largely overlooked the direct effects of household responsibilities on earnings, instead focusing attention on the effect of differences in household roles on human capital accumulation. According to human capital theory, women who bear the majority of household and child care responsibilities may expect discontinuous labor force participation and fewer total years in the labor force than men. Thus women will have fewer years over which to reap the rewards, and hence will optimally choose to acquire less human capital. Further, employers will provide less specific training to women workers in anticipation of their higher turnover.

By this argument, the lower average earnings of women are attributable to lower average quantities of human capital. Yet women are invariably found to earn less than men with equivalent human capital characteristics. Further, as the labor force participation rates of men and women continue to converge, differences in human capital will decrease in importance as an explanation of wage differences between men and women. Yet most women, even those with market jobs, continue to assume the

primary responsibility for household chores. The purpose of this paper is to examine the direct effect on market productivity of the dual responsibilities of market and nonmarket work.

I. Data Set and Empirical Results

To investigate the direct role of housework in affecting wages, I use data from the 1987 *Panel Study of Income Dynamics (PSID)* (Wave 20). This data set has information on 7061 households. Each head of household was asked to respond to a series of questions about sources and amounts of income, labor market activity, and personal background, and to answer a parallel set of questions about their spouse, if the head is married.

For the purposes of this study, the advantage of this data set is that it contains a measure of time spent on housework. Heads of households were asked to answer the following question, for themselves and for their spouse: “about how much time do you (does your spouse) spend on housework in an average week? I mean time spent cooking, cleaning, and doing other work around the house.”

The average values of time spent on housework for the sample members that are employed or temporarily laid off are reported in Table 1. The values are reported by gender, marital status, and presence of children under age 18 in the housing unit. As one would expect, the average values of time spent on housework for parents of either gender exceed the corresponding values for individuals without children under age 18.¹

¹It should be noted that the average values reported in this survey are far below those reported in response to a similar question in the 1977 *QES*, as well as in a data set I collected in 1986 (see my forthcoming article). Possible reasons for these relatively low values include that all values were reported by the head of household for both the head and spouse, which may lead to

*Department of Economics, University of Wyoming, Laramie, WY 82071.

TABLE 1—AVERAGE VALUES OF TIME SPENT
PER WEEK ON HOUSEWORK
BY MARKET WAGE EARNERS IN SAMPLE

	Mean ^a	Sample Size
Married Men:		
Children	8.96 (7.58)	1217
No Children	7.58 (6.29)	532
Married Women:		
Children	19.42 (10.61)	1031
No Children	15.16 (8.21)	587
Not-Married Men:		
Children	10.43 (7.50)	70
No Children	7.73 (5.75)	395
Not-Married Women:		
Children	13.98 (8.32)	300
No Children	10.06 (7.36)	386

^aStandard deviations are shown in parentheses.

Most noteworthy are the dramatic differences in housework time by gender. The time spent on housework by women exceeds that of their male counterpart in every category, statistically significant at the .01 level or better in every case. Married women with children average nearly 20 hours per week on housework, more than double that of married men with children. Married women without children average 15 hours of housework per week, which is 7.58 hours more per week than married men without children. Note that the time spent on housework for married women is usually reported by their husbands, and may be an underestimate of the true value.

Wages and time spent on housework are determined jointly, with higher wages mak-

ing substitutes for nonmarket work more affordable. This suggests that the appropriate procedure is to estimate a two-equation wage-housework system. Accordingly, Table 2 presents the results of this estimation, where the estimation procedure is two-stage least squares (2SLS).

The wage equation takes the standard human capital specification, augmented by hours of time spent on housework. The log of hourly wage is regressed on hours per week spent on housework, years of education, the log of years of tenure with employer, and years of full-time work experience and its square. In addition, the regression includes dummy variables equal to one if the worker is white, handicapped, married, in a job covered by a union contract, in a white-collar job, is employed full time, or resides in the South. The inverse Mill's ratio, calculated from the full sample of workers and nonworkers, is also included in the wage equation to correct for possible selection bias that may occur since we observe only wages of individuals whose market wage exceeds their reservation wage.²

One determinant of time spent on housework is the market wage rate, since higher market wages make substitutes for own housework more affordable. In addition, time spent on housework will be affected by a variety of individual and household characteristics, as well as by cultural differences and individual attitudes. The household characteristics include number of children (in four different age ranges), number of rooms in the home, and dummy variables equal to one if the individual is married and if the individual lives in a house (rather than an apartment or trailer). Differences in attitudes or cultural differences may be accounted for by race, education, age, and

underreporting of wives' time spent on housework. In addition, unlike the QES and my survey that requested information on time spent daily on household chores including yard work, repairs, and shopping, the PSID question only requested a summary measure for the week and did not prompt for household chores other than cooking and cleaning, again leading to the likelihood that the time spent on housework is underreported.

²The inverse Mill's ratio was estimated from a probit equation (not reported) that estimated the probability of labor force participation from age, race, handicapped status, marital status, number of children under age 6, number of children between ages 6 and 18, a dummy variable indicating that there were no children under age 18, years of education, and the wage rate for unskilled workers in the county of residence.

TABLE 2—2SLS ESTIMATES OF WAGE AND HOUSEWORK EQUATIONS^a

Independent Variables	Men		Women	
	(1)	(2)	(1)	(2)
<i>Housework</i>	.036 (.012)		-.006 (.003)	
<i>Log(Wage)</i>		-.047 (1.004)		-7.002 (1.812)
<i>Education</i>	.063 (.005)	-.121 (.090)	.063 (.005)	-.216 (.135)
<i>Log(Tenure)</i>	.085 (.009)		.055 (.010)	
<i>Experience</i>	.038 (.013)		.070 (.017)	
<i>Experience Squared</i>	-.0007 (.0004)		-.002 (.0005)	
<i>White</i>	.218 (.027)	-1.057 (.392)	.095 (.018)	.961 (.478)
<i>Handicapped</i>	-.153 (.069)		-.030 (.033)	
<i>Union</i>	.146 (.025)		.190 (.022)	
<i>White Collar</i>	.188 (.025)		.158 (.020)	
<i>Full Time</i>	.200 (.045)		.095 (.025)	
<i>South</i>	-.083 (.021)	.455 (.336)	-.082 (.018)	-.005 (.441)
<i>Married</i>		-.204 (.476)		3.896 (.590)
<i>Age</i>		-.084 (.361)		2.196 (.734)
<i>Age Squared</i>		.001 (.004)		-.026 (.009)
<i>Children under 2</i>		1.739 (.467)		2.940 (.687)
<i>Children age 3-5</i>		1.519 (.461)		3.644 (.604)
<i>Children age 6-13</i>		1.274 (.415)		.685 (.469)
<i>Children age 14-17</i>		.964 (.548)		-.219 (.666)
<i>Other Family Income</i>		.00002 (.00001)		.00002 (.00001)
<i>Family Size</i>		-.629 (.301)		-.242 (.385)
<i>House</i>		-.460 (.384)		-.062 (.506)
<i>No. of Rooms</i>		.265 (.113)		.529 (.142)
<i>Unskilled Wage Rate in County</i>		.049 (.150)		.182 (.195)
<i>Mill's Ratio</i>	.068 (.163)		-.028 (.047)	
<i>Intercept</i>	.308 (.127)	11.269 (5.913)	.553 (.079)	-18.356 (11.258)
<i>Adjusted R²</i>	.38	.02	.41	.18

Note: Col. (1) is Log(Wage); Col. (2) is Housework.

^aStandard errors are shown in parentheses.

residence in the South. Other family income (net of the individual's own labor income) is included to indicate the household's ability to afford substitutes for nonmarket work. The county wage rate for unskilled labor proxies for the cost of substitutes such as paid housekeepers.

The results of the simultaneous estimation of the wage-housework system are presented in Table 2. The wage equation conforms to standard estimates of wage equations. Wages rise with tenure and years of work experience at decreasing rates, more educated workers tend to earn higher wages, as do union, white-collar, full-time, and white workers, and workers not located in the South.

The time spent on housework by both men and women is primarily affected by the presence of children and by the number of rooms in the home. White women and married women spend more time on housework, while white men and men in larger families spend less time on housework. The negative effect of family size on housework performed by men, after controlling for the number and ages of children, may be due to the presence of other adults (for example, parents) that can help with housework. Other family income does not have a significant effect on time spent on housework for either men or women.

Most noteworthy is the significantly negative effect of housework on wages, and of wages on housework, for the sample of women. Each extra hour of housework reduces women's hourly wages by an average of .6 percent, while each extra dollar per hour in wages earned by women reduces her time spent on housework by about 2.5 hours. Surprisingly, men's wages are positively and significantly related to time spent on housework, while time spent by men on housework is not affected by their wage.

II. Discussion

The results indicate that women's wages, but not men's, are reduced by time spent on housework. Further, the time spent by women on housework is inversely related to her own earnings, but is not affected by the household's other income. Men's time on

housework is unaffected by their wage or by other family income.

The basis for the inverse relation between women's wages and time spent on housework may be due to a number of related factors. These include the possible direct effect of housework on market effort and the possibility that household roles affect demand for working conditions and thereby wages as a compensating differential. A further possibility is that women who have demonstrated that family life is a priority are placed on a so-called "mommy track" with reduced work responsibilities and promotion prospects.

Housework may have a direct effect on market productivity, after controlling for any effects that anticipated household responsibilities have on human capital acquisition. The direct effect may be caused by a reduction in the amount of effort available for market work. This result is consistent with my earlier research, using two other data sets, that indicates that housework has a negative effect on women's wages (see my 1985 and forthcoming articles). Yet time spent on housework apparently has a positive effect on the wages of men.

Since women spend more time on housework than men, any negative effect of housework on wages may begin at a point beyond the average amount of time spent on housework by men. Further, it is more likely that the timing of household chores is different for men and women. Women are more likely to take responsibility for chores that have a time element associated with them, such as cooking a meal or arranging doctors' appointments for children. Because such activities make schedules less flexible, market work is more likely to be disrupted for women than men. For instance, women may be less likely to work late than men, and more likely to take time off work to make and meet family-related appointments.

To the extent that housework interferes with women's market work because of scheduling and physical and mental effort, we would expect to see women in jobs with characteristics that reflect these different requirements. If these job characteristics warrant lower pay as a compensating dif-

ferential for favorable job conditions, then the negative effect of housework on wages may be spurious, and instead due to the correlation of housework with unobserved working conditions that warrant lower wages as a compensating differential. This hypothesis cannot be tested using the *PSID* data set, but I have tested this theory using data I collected, with the results presented in my forthcoming article. I find that although men and women are in jobs with very different characteristics, time allocated to household responsibilities has an independent negative effect on women's wages, controlling for differences in working conditions and human capital.

A second hypothesis regarding the manner in which housework reduces wages may be due to the indirect effect referred to as the mommy track. The notion of a mommy track suggests that family responsibilities and careers are fundamentally incompatible, and the expectations on the job for women should accordingly be scaled down. Women that have demonstrated that they are taking on household chores may be de facto placing themselves on a slower track with respect to promotions. Thus lower wages accompanying greater household responsibilities may be caused by women being promoted at different rates than men with otherwise similar human capital characteristics.

However, despite the popular press notion that women are victims of men and are forced to do housework to the detriment of their careers, it is worth noting that even childless unmarried women spend more time on housework than their male counterpart. This suggests that at least some of the extra time on housework spent by women is due to differences in tastes.

REFERENCES

- Hersch, Joni, "The Effect of Housework on Earnings of Husbands and Wives," *Social Science Quarterly*, March 1985, 66, 210-17.
- _____, "Male-Female Differences in Hourly Wages: The Role of Human Capital, Working Conditions, and Housework," *Industrial and Labor Relations Review*, forthcoming, July 1991.

Gender Differences in Labor Market Effects of Alcoholism

By JOHN MULLAHY AND JODY L. SINDELAR*

Little is known about the role of specific health problems in affecting labor market productivity. Even less is known about gender differences in the labor market effects of such health problems. Current knowledge of health effects is based largely on samples composed exclusively of men, a common practice in both economics and health research. In the latter case, even a congressional mandate to incorporate females in study samples is reputed to have had little effect (Patricia Schroeder, 1990).

In this study, we attempt to determine the structure of gender differences in labor market responses to alcoholism. We use a relatively new data source that allows such comparisons by gender in a large community-based sample. Previous studies have established that there are significant gender differences in labor market behavior. Differences in prevalence rates of alcoholism by gender are also well established. It is estimated that 3 percent of females are currently suffering from alcoholism and twice that many have exhibited symptoms at some time; for males the numbers are 10 and 20 percent, respectively. There is some medical evidence to indicate that physiologically, women and men respond differently to alcohol. For example, a recent study suggests that women have greater vulnerability to the acute and chronic health conditions associated with alcoholism (Mario Frezza et al., 1990).

I. Background

There are many debates surrounding alcoholism and alcohol consumption, ranging

from the effectiveness of treatment of alcoholism to the productivity effects of alcohol. There is even debate as to the sign of the effect of alcohol and alcoholism on the earnings of men. From Adam Smith, to Irving Fisher's denunciation of drinking as a detriment to productivity, to more recent studies, many have supported the idea that alcoholism or alcohol consumption has a negative effect on income and productivity. Yet others have found either insignificant effects (Lee Benham and Alexandra Benham, 1982) or even positive effects (Mark Berger and J. Paul Leigh, 1988). There is also a debate over whether alcoholism is a disease or a result of rational choices. At least in the medical literature, alcoholism is generally considered a "disease," "disorder," or "allergy."

Another cloudy area is the causal path by which alcoholism affects an individual's labor market outcomes. Early onset of alcoholism may retard educational achievement (see our 1989 and 1990b articles) and affect marital history. Current alcoholism could affect labor supply decisions (by changing the value of time at home vs. in the market), wage rates (by affecting occupational choice and reliability of the worker), nonwage income (by affecting marital status, spouse's labor supply decisions and/or transfer payments), and accumulated wealth (by influencing savings rates, past wage rates and labor supply, earned pension rights, etc.). Thus alcoholism could have direct effects on earnings as well as indirect ones. While we can specify neither the causal relationships nor their lifetime paths, it is important to keep these in mind when viewing and interpreting the data.

II. The Data

We use multiple site data from the Epidemiologic Catchment Area (ECA) survey. This set of surveys was designed to assess the prevalence of psychiatric problems (including alcoholism) by socioeconomic and

*Department of Economics, Trinity College, Hartford, CT 06106, and Resources for the Future; and Department of Epidemiology and Public Health, and Institution for Social and Policy Studies, Yale University, New Haven, CT 06510, respectively. Grant R01AA08394 from NIAAA supported this work. We thank Bernie Devlin for his programming assistance and Jean Mitchell for comments.

demographic characteristics in a community setting (see D. A. Reiger et al., 1984). Prior to the availability of the ECA surveys, there were no large samples assessing alcoholism and other disorders that contained professional measures of disorders. The ECA study comprised five sites. However, as one site gathered no information on income, we confine our analysis to data from Baltimore, MD, Durham, NC, Los Angeles, CA, and New Haven, CT.

We use as our definition of alcoholism whether the individual has ever met the criteria for alcohol dependence or abuse (see our 1990c paper for a discussion of the measurement of alcoholism). We refer to this as *ALCEVER*. We focus on *ALCEVER* as it is more likely to be exogenous to current behavior than is whether the person is currently manifesting symptoms. We treat *ALCEVER* as an exogenous variable because this is consistent with the disease perspective of alcoholism, and because it is econometrically necessary as there are no reasonable instrumental variables in the data set. This is also in keeping with the standard practice of treating health as an exogenous variable in labor market regressions.

To obtain an overview of gender differences, we compare the mean values of various socioeconomic characteristics of alcoholics to nonalcoholics for men and women separately. For both genders, nonalcoholics have higher educational attainment, are more likely to be working, are more likely to have white-collar jobs, and have higher incomes but are less likely to have transfer income.

Alcoholism appears to affect women somewhat differently than it affects men. For example, the percentage difference in household income by alcoholic status is bigger for women than for men. Further, alcoholic women are more likely than alcoholic men to never marry, have fewer children, and have more psychiatric disorders.

III. Life Cycle Dimensions to Alcoholism Problems

In previous work (1990c), we hypothesized that the labor market effects of alco-

holism would vary over the life cycle. That is, in the earlier ages, those suffering from alcoholism might either pay less attention to their schooling or drop out completely. In either case, young alcoholics may have higher wages due to more labor market experience and may also work more hours than their nonalcoholic counterparts. In the prime working ages, alcoholics would lose their lead in earnings and would have relatively lower wages as a consequence of their lower educational attainment, *ceteris paribus*. Yet, older alcoholics might tend to remain in the labor force longer than their nonalcoholic counterparts who have begun to withdraw from the labor market due to their greater accumulation of wealth and pension rights.

In work that used a sample of males from the New Haven site of the ECA survey (our 1990c paper), we found some evidence consistent with these hypotheses. Here, with a more comprehensive sample, we analyze how such relationships may differ by gender examining how alcoholism relates to work status (presently working full time), household income, and personal income. By way of a simple comparison of means, Table 1 provides some insights into these relationships by age groups. These comparisons are inherently not *ceteris paribus* comparisons; this issue we address below in Section IV.

In the three older age groups, the probability of labor market participation for both males and females is reduced substantially when individuals are alcoholic. This is in contrast to the youngest cohort where no negative effects are seen. As shown by the Household Income and Personal Income (displayed for workers only) panels of Table 1, there is little ambiguity about the relationship between alcoholism and income for the prime age groups 30–59. For both genders and for both household and personal income, alcoholism results in negative effects except for personal income of females ages 30–44.

Consistent with our earlier work with the New Haven sample, working alcoholic males in the youngest age category have somewhat greater income than nonalcoholics in the same age range, although this does not hold for females of the same age. Among work-

TABLE 1—SAMPLE MEANS BY AGE AND ALCOHOLISM STATUS: WORK STATUS; HOUSEHOLD INCOME; AND PERSONAL INCOME FOR WORKING FEMALES AND MALES

Ages	Females <i>ALCEVER</i>		Males <i>ALCEVER</i>	
	= 0	= 1	= 0	= 1
Work Status (Working = 1)				
18–29	.50	.50	.72	.75
30–44	.59	.48	.89	.79
45–59	.49	.26	.79	.66
60–64	.30	.24	.50	.37
Household Income (in \$1,000)				
18–29	22.0	18.6	22.8	23.9
30–44	24.7	20.8	28.7	25.6
45–59	23.1	22.8	30.1	27.3
60–64	16.7	21.1	28.7	21.9
Personal Income (in \$1,000)				
18–29	10.4	10.0	14.2	14.7
30–44	14.3	15.3	23.4	20.8
45–59	13.2	11.9	24.7	21.5
60–64	12.1	13.4	24.7	19.3

ing women, however, the older alcoholics (ages 60–64) have larger mean incomes than their nonalcoholic counterparts (although the sample size on which this comparison is based is small). A striking result revealed by these age-income profiles is that household income plummets for nonalcoholic but not for alcoholic working women in the oldest group. Consistent with this is the fact that personal income for alcoholic working women increases in moving to the oldest age group while it declines for nonalcoholic women.

Table 1 suggests, among other things, that the two middle-age groups for both sexes are relatively homogeneous with respect to the labor market effects of alcoholism, while both the youngest and the oldest group are somewhat different. We thus proceed to focus on the prime-aged working group (ages 30–59) on the premise that they and only they can be reasonably pooled together.

IV. Econometric Models

Given these relatively homogenous samples of prime-aged men and women, we attempt to address the *ceteris paribus* issue

raised above. That is, although we have shown that prime-aged alcoholics tend on average to work less and have lower incomes when working, it is not clear from the comparisons in Table 1 whether this is due to alcoholism per se, or rather to other characteristics of alcoholics.

To address this issue, we estimate several models of the probability of working, household income, and personal income, where income is measured in logs of interval mid-points. For the income equations, we estimate models for both the entire sample and also for the presently working subsample. Personal income, especially for workers, is closer to a measure of the individual's labor market productivity. However, when analyzing household income, one captures the possible effects of marital status and also of spouse's labor supply and earnings responses. Spouses may adjust their labor supply in response to a spouse's alcoholism for a variety of reasons, including the income effect and a change in the pleasantness vs. the productivity of time at home. The possibilities of alcoholism-related changes in the marital status and of assortative mating also complicate the issue.

The models are specified as functions of alcoholism as well as other more-or-less standard covariates that typically appear in earnings functions. Table 2 describes the variables in addition to *ALCEVER* that are used in the models.

The idea is to specify first a model in which the alcoholism variable is left to absorb the effects of the other human capital variables that are not included but that may be correlated with alcoholism. In a sense, these results mimic those in Table 1 except here we have adjustments for age, race, and the ECA survey site. This coefficient on alcoholism is an estimate of the total effect of alcoholism. Then we add variables that may be correlated with, and perhaps to some degree determined by, alcoholism. (Unfortunately, our data do not allow us to address satisfactorily the cause-effect issue.) In this latter case, the coefficient on alcoholism captures only part of the total effect, the rest of the effect occurs through other human capital components that vary with alcoholism. Thus there are two possible mea-

TABLE 2—REGRESSION COEFFICIENTS ON *ALCEVER*
BY GENDER AND WORK STATUS: EFFECTS OF
ADDITIONAL HUMAN CAPITAL COVARIATES
(asymptotic *t*-ratios in parentheses)

	Work Probability ^a			
Sample	A	B		
All Individuals				
Females	-.79	-.67		
(N = 2795)	(4.53)	(3.01)		
	[-.19]	[-.16]		
Males	-.87	-.51		
(N = 2162)	(6.59)	(2.89)		
	[-.11]	[-.07]		
	Income ^b			
	Household		Personal	
Sample	A	B	A	B
All Individuals				
Females	-.47	-.08	-.16	-.11
(N = 2778)	(7.47)	(1.47)	(1.60)	(1.12)
	{-.38}	{-.08}	{-.15}	{-.10}
Males	-.20	-.03	-.26	-.10
(N = 2140)	(5.27)	(0.81)	(6.50)	(2.45)
	{-.18}	{-.03}	{-.23}	{-.10}
Workers				
Females	-.28	-.03	-.10	-.03
(N = 1596)	(3.50)	(0.45)	(1.29)	(0.39)
	{-.24}	{-.03}	{-.10}	{-.03}
Males	-.09	-.001	-.12	-.03
(N = 1823)	(2.76)	(0.05)	(3.44)	(1.01)
	{-.09}	{-.001}	{-.11}	{-.03}

Note: Col. A. Controls for: geographic site, race, age, and age squared. B. Adds: education, education squared, number of children, marital status, transfer income, other household income (in the personal income equation only), and other disorder measures whose inclusion varies by gender: for females and males, anxiety, drug problems, and an aggregate of any other disorder; for males only, antisocial personality; and for females only, depression.

^aEstimated by logit. Figures in square brackets are $\hat{\beta}_j \hat{\pi}(1 - \hat{\pi})$, i.e., the "partial derivatives" of the predicted probabilities with respect to the dummy.

^bFigures in curly brackets are the estimated percent changes in income due to turning on *ALCEVER*, i.e., $\exp(\hat{\beta}_j) - 1$.

tures of effects of alcoholism. In one, the coefficient on alcoholism absorbs all of the effects (direct and indirect), while in the other, the coefficient estimates the partial effect holding other variables constant.

The regressions that include only the sparse set of covariates (cols. A in Table 2) provide an estimate of what we term the total effect of alcoholism on labor market

success. These results are striking and indicate that in what we feel to be a relatively homogeneous sample, alcoholics have dampened labor market success when compared with nonalcoholics.

For the logit models of presently working, the *ALCEVER* coefficients are negative and statistically significant for both genders. While the point estimate itself is larger for males, the partial derivative of the probability of working with respect to *ALCEVER* (displayed in brackets) is almost twice as large for females as for males. The effects are similarly negative and significant (except for personal income for females) for the log-income models (cols. A). The percentage reductions in income implied by the point estimates of the *ALCEVER* coefficient (displayed in curly brackets) indicate that these effects are sizable. We consider comparisons of household income for all individuals and personal income for workers only as the most meaningful comparisons; this is because the former will capture the effects of alcoholism on household structure and labor supply while the latter is more likely to capture individual productivity effects. These results indicate that the effects of alcoholism are greater on household income for women, but are similar by gender for personal income.

When we include the set of human capital variables omitted from the first estimates (cols. B), the coefficients on alcoholism are reduced in magnitude and significance for both sexes and for all measures of labor market success. This suggests a strong relationship between alcoholism and these other variables, and, therefore, supports the possibility of important indirect effects. In the context of omitted variable bias, this is hardly surprising. Note that the effects of alcoholism on labor market participation are still significant even when controlling for these other determinants, with the effect for females again estimated to be about twice that of males. However, for income, these effects generally become insignificant when moving to column B. Thus the effect of alcoholism is found to be stronger on labor force participation than it is on income, especially personal income.

V. Summary

Access to the multiple site ECA data allows the first large population-based inquiry into gender differences in the effects of alcoholism on labor force participation and earnings. In summary, we find that alcoholism typically has negative effects on both labor force participation and income for our sample. However, the effects vary across the life cycle and by gender. The significance depends on what variables one controls for and whether one is examining participation or income. The range of the results that we present provides relevant bounds of the magnitude of the effects of alcoholism on these prime-aged individuals. These results help to explain how studies could produce different estimates of the effects of alcoholism on income. The effect can depend on what variables one controls for, the age distribution of the sample, and whether one is examining income or participation.

REFERENCES

- Benham, Lee and Benham, Alexandra**, "Employment, Earnings, and Psychiatric Diagnosis," in V. Fuchs, ed., *Economic Aspects of Health*, Chicago: University of Chicago Press, 1982.
- Berger, Mark C. and Leigh, J. Paul**, "The Effect of Alcohol Use on Wages," *Applied Economics*, October 1988, 20, 1343-51.
- Frezza, Mario et al.**, "High Blood Alcohol Levels in Women," *New England Journal of Medicine*, January 1990, 322-2, 95-99.
- Mullahy, John and Sindelar, Jody**, "Lifecycle Effects of Alcoholism on Education, Earnings, and Occupation," *Inquiry*, Summer 1989, 26, 272-82.
- ____ and _____, (1990a) "Gender Differences in the Effects of Mental Health on Labor Force Participation," in I. Sirageldin, ed., *Research in Human Capital and Development: Female Labor Force Participation*, Greenwich: JAI Press, 1990.
- ____ and _____, (1990b) "An Ounce of Prevention: Productive Remedies for Alcoholism," *Journal of Policy Analysis and Management*, Winter 1990, 9, 249-53.
- ____ and _____, (1990c) "Alcoholism, Work, and Income Over the Life Cycle," presented at AEA annual meeting, Washington, D.C. 1990.
- Reiger, D. A. et al.**, "The NIMH Epidemiologic Catchment Area (ECA) Program: Historical Context, Major Objectives, and Study Population Characteristics," *Archives of General Psychiatry*, October 1984, 41, 934-41.
- Schroeder, Patricia**, "Pay Attention to Women's Health," *Journal of NIH Research*, August 1990, 2, 38.

Europe Post-1992: Internal and External Liberalization

By ALEXIS JACQUEMIN AND ANDRÉ SAPIR*

The gradual completion of the single European Community (EC) market is a process of internal liberalization that takes place within the framework of a worldwide dynamics of structural adjustment. At the *internal level*, the process is well on its way for at least two reasons. First, most of the regulations needed to remove the remaining physical, technical and fiscal barriers between member states have been adopted. Second, corporate strategies have largely anticipated the conditions of the post-1992 single market through various forms of restructuring, including a growing concentration on the main product lines, an extension of geographic coverage, and a multiplication of cooperative arrangements, mergers, and acquisitions. At the *external level*, foreign competitors supplying products for which demand is income elastic will benefit from the European income growth induced by market integration. They will also benefit from scale economies provided by a large integrated market in which they compete with EC firms on equal terms. It remains true that many nontariff barriers affect external policies for products and services, and that the dangers of an EC turning more inward cannot be excluded.

The paper argues that at the micro and macro level, a combined internal and external liberalization is necessary for the full achievement of the post-1992 gains. This requires transnational rules of the game sustaining extra- as well as intra-EC competition, and cooperative policies making compatible regionalism and multilateralism.

*Université Catholique de Louvain and Université Libre de Bruxelles, Belgium, respectively. Comments by P. Bolton, C. Hamilton, and P. Messerlin are gratefully acknowledged.

I. Internal and External Competition

Theoretical and empirical research suggest that import competition within European markets imposes a major constraint on domestic firms' price-cost margins. The program for the completion of the EC's internal market by 1992 is largely based on the effects expected from a reinforcement of such a constraint. The 1992 program requires the removal of barriers still affecting intra-EC trade, and hence the strengthening of European competition. According to the European Commission's assessment of the economic effects of this liberalization, the overall result will be a significant welfare gain. Given these expectations, the combination of internal and external liberalization may be a superior policy to internal liberalization alone if external liberalization also exerts an appreciable competitive impact.

Two recent studies shed light on this question. The first (our 1990 paper) investigates empirically the relative trade discipline of intra- and extra-EC imports on European industry performance. Our estimates reveal that only extra-EC imports exert a significant disciplinary effect on price-cost margins. A second study (by Damien Neven and Lars Röller, 1990) concludes that the elimination of intra-EC barriers should increase extra-EC imports more than intra-EC imports. Hence, there is evidence that the main competitive pressure, both actual and potential, comes from the rest of the world rather than from within the EC.

More specifically, the external impact of 1992 will be felt most in those sectors where the EC's competitive position has weakened most since the mid-1970's (see Sapir, 1989). Those sectors can be divided into two clearly

distinct categories. The first consists of industries that are characterized by a low or average level of R&D, and by slight or average economies of scale: shipbuilding, footwear, textiles, and clothing. Despite considerable external barriers, European producers have suffered appreciable market losses in these sectors, generally owing to the comparative advantage enjoyed by the developing countries. The second category consists of high-tech sectors characterized by relatively substantial economies of scale, important learning processes, and high sunk costs: telecommunications, electronics, and computers. In these sectors the weak performance of EC producers contrasts with the gains by United States and Japanese exporters.

In both these sector categories, important choices will have to be made in the run-up to 1992. In the case of the low- or average-technology sectors, steps must be taken to ensure that EC commercial policy is not used in place of a more effective structural adjustment policy. In the high-tech oligopolistic sectors, the public authorities will have to resist purely defensive reactions and calls for protection, and will have to ensure that policies designed to promote competitiveness and competition are put in place.

Competition policy is one of the most important guarantees for the internal and external liberalization of the Common Market. It is by no means certain that, post-1992, economic agents will accept the operation of competition. As experience following the lowering of tariff barriers has shown, the EC authorities may well be confronted with growing private and public strategies that seek to diminish or distort competition. The EC authorities must then ensure implementation of *credible rules* that are directly applicable to all, including third-country companies. The competition rules of the Treaty of Rome are already applicable to both public and private restrictions of competition and, in both cases, have been tightened. Concerning public restrictions, an example relevant for the external impact of 1992 is the increased transparency of state-

aid policy, required by the Commission and, when necessary, the recovery of aids granted illegally.

Concerning private restrictions, a central regulation about mergers and acquisitions was adopted in December 1989. This regulation sets up EC controls over community-wide, cross-border operations, and there is a mandatory prior notification of the planned mergers of this kind. One characteristic is also crucial for the external impact of 1992: for assessing whether a merger is or is not compatible with the Common Market, the only basis is its impact on effective competition, at the exclusion of cost savings or other offsetting efficiencies. This absence of an "efficiency defense" in the text of the regulation reduces sharply the dangers of mixing competition policy with industrial policy (see Jacquemin, 1990).

Ambiguity remains over the interactions between EC trade and competition policy. Compared to the rule-based competition policy (probably the most impersonal and the least discriminatory means of social control of an economy), trade policy allows much more discretionary power. According to Article 113 of the Treaty of Rome, "the common commercial policy shall be based on uniform principles," but the trade rules do not exclude the adoption by the Council of Ministers of anticompetitive protectionist measures. In fact, maintaining intra-EC competition is a much more accepted goal than safeguarding competition from outside.

More specifically, the pressures for not disrupting economic transactions between member states could lead to a difficult choice between the erosion of intra-EC free trade and the erosion of free trade with the rest of the world. One unhappy implication is that in several domains, not only trade and competition policy follow different roads, but also interact in a perverse way. While a free-trade policy is part of competition policy, the use of instruments such as antidumping duties and voluntary export restraints not only affects extra-EC trade but, through a feedback effect, mitigates competition within the EC. Some authors (for

example, Patrick Messerlin, 1990) argue that in fact European antidumping (allowed by Article VI of GATT) actions have a procartel and promerger propensity. Such actions increase the capacity to cartelize for firms unable to collude without some kind of public support; they also induce EC firms to merge with foreign firms operating in EC markets to the extent that EC-owned foreign exporters are more immune to antidumping measures.

At a time of globalization of business strategies, the solution is not simply to improve antidumping criteria and procedures, but also to develop international cooperative agreements between antitrust authorities, intended to achieve consensus concerning principles and implementation of a world competition policy. Progress is on the way through multilateral discussions (OECD, UNCTC), but also EC bilateral negotiations (with U.S. Federal Trade Commission and Department of Justice; Japanese MITI and Fair Trade Commission).

The main point here is that the distinction between "internal" and "external" competition is more and more blurred given the transcontinental nature of direct investments and the importance of intrafirm trade, the increasing mixing up of the local and international content of final products, the continuous process of delocation of activities all over the world. In such a setting, a strong EC domestic competition policy, coupled with a relatively protective external trade policy, will not only limit the benefits of competition, it will be unable to control effectively the multiplication of restrictive and monopolistic practices in world markets. It is necessary to consider these questions as transnational issues that cannot be solved by unilateral "beggar thy neighbor" initiatives, but will instead require worldwide agreements applicable to "free-rider" countries.

II. Regionalism and/or Multilateralism

Trade economists have long acknowledged that the spatial location of nations plays an important role in shaping trading patterns. (The role of spatial location in

international trade has recently been given a new twist by Paul Krugman, 1990, in the context of increasing returns to scale.) The notion that proximity matters in international trade led European researchers to develop the so-called "gravity model" in the early 1960's. This model postulates that the intensity of bilateral trade flows between two nations is inversely related to the distance between them. Empirical applications of the model have generally defined proximity in a broad economic sense, including geographical, cultural, and institutional dimensions.

Despite their lack of precise theoretical foundation, gravity models have been extremely successful empirically. The natural tendency for countries to trade intensely with their neighbors has long been invoked as a legitimate motive for regional integration. Thus, although the most-favored-nation (MFN) principle was already central, at least as far back as the early nineteenth century, many commercial pacts in Europe allowed departures from MFN obligations on the grounds of either "customs union treaties" or "regional clauses." The latter generally reflected "close ties of sentiment and interest arising out of ethnological, or cultural, or historical political affiliations. *Proximity* was usually a characteristic of the countries so related..." (Jacob Viner, 1950, pp. 18-19, emphasis added).

After World War II, these two exemptions from MFN obligations were enshrined in the Havana Charter for an International Trade Organization (ITO). The Charter recognized "the desirability... of closer integration... [through] the formation of a custom union or a free-trade area" (Article 44). It also authorized the creation of "preferential agreements for economic development and reconstruction" provided certain conditions were fulfilled. First among these conditions was the requirement that "the territories of the parties to the agreement are *contiguous* one with another, or all parties belong to the *same economic region*" (Article 15, emphasis added).

Although the Havana Charter was never implemented, the possibility to establish a customs union or a free-trade area was con-

firmed by the General Agreement on Tariffs and Trade (GATT), provided such arrangement meets a double test: it applies to "substantially all" trade between the partner nations, and it does not raise the level of trade barriers against third nations (Article XXIV). By contrast, GATT did not carry over the provisions of the Charter with respect to preferential agreements among parties that are either "contiguous" or belong to the "same economic region." In the eyes of the founding fathers of GATT, the complete elimination of tariffs among a group of nations, regardless of their geographical location (as in a customs union or a free-trade area) was clearly to be preferred to the partial elimination of tariffs, even among nations belonging to the same region (as in preferential agreements).

This does not mean, however, that the notion of propinquity has had no relevance for GATT-sanctioned customs unions or free-trade areas. On the contrary, the proximity between parties to such arrangements is generally regarded by economists as an important factor in judging their desirability. Thus Paul Wonnacott and Mark Lutz argue that: "Trade creation is likely to be great, and trade diversion small, if the prospective members of an FTA [a free-trade area] are *natural trading partners*." This is more likely to be case if "the prospective members [are] *close geographically*" (1989, p. 69, emphasis added). Thus, a regional integration arrangement that involves the elimination of barriers on "substantially all trade" among nations in close geographic proximity is likely to reinforce natural trading patterns and benefit world trade. As such, it can be referred to as a "natural integration."

The EC fits perfectly well the description of "natural integration," at least as far as manufactured goods are concerned; however, the Common Agricultural Policy (CAP) has modified "natural" trade patterns and resulted in trade diversion. This is confirmed by empirical estimations by Bela Balassa and Luc Bauwens that indicate that the EC has had far less impact on the pattern of international trade than "natural" proximity factors such as geographical dis-

tance and cultural similarity (see their 1989 paper, Table 1). The same should apply to Europe 1992 and the future integration arrangements between the EC and European Free Trade Area countries and between the EC and the countries of Eastern and Central Europe. The key is the external trade policy. These arrangements will be beneficial to world trade provided their external trade policy is liberal.

Until recently, the United States was the champion of GATT and the multilateral system. However, political, economic, and intellectual developments have combined to bring about a new U.S. trade strategy. The "diminished giant syndrome" has brought distrust by the United States in the ability of the multilateral system to serve its interests (see Jagdish Bhagwati, 1990). A strong consensus has emerged in this country in favor of an aggressive, result-oriented trade policy. The new rhetoric did not take long to be implemented. Bilateral actions have multiplied, free-trade arrangements have been signed, and more steps are under way according to a logic of "U.S. trade bloc strategy" (see Rudiger Dornbusch, 1990). Such development, combined with the deepening and widening of integration in Europe and possible regional actions by Japan, has created a sentiment of fragmentation of the trading system into three blocs.

The emergence of a tripolar world puts the future of regionalism at the crossroads. In the years ahead, regionalism can either contribute or be detrimental to the multilateral trading system. The two scenarios are possible, and correspond to a distinction between "natural integration" and "strategic integration." If the three blocs decide to play a noncooperative game, natural integration will be used as a leverage for strategic integration, reflecting the fact that the evolution of integration is not primarily influenced by the *exogenous* trade pattern, but also by *endogenous* trade and industrial policies. (See Jacquemin, 1987, for a similar distinction in industrial organization.) Such a scenario would lead to costly trade wars that may be beneficial to certain nations at the expense of others and/or collectively disadvantageous. This danger is especially

important in activities with large economies of scale and few producers, where strategic policy interventions are already frequent in the EC, the United States, and Japan. On the other hand, a favorable outcome (the compatibility of regionalism and multilateralism) may be possible, provided the three major actors play a cooperative game both among themselves *and* with other nations.

In conclusion, the interconnected internal and external liberalization of the European market, post-1992, can lead to substantial welfare gains. In order to obtain such a global liberalization, credible worldwide agreements about transnational rules of the game are required. Existing GATT rules and disciplines have stood the test of time, but need to be reinforced. It is equally important to recognize that present GATT rules deal only with government interventions. They are silent on everincreasing restrictive business practices. The time has come to enlarge the scope of GATT and consider the establishment of worldwide competition rules.

REFERENCES

- Balassa, Bela and Bauwens, Luc, "The Determinants of Intra-European Trade in Manufactured Goods," in A. Jacquemin and A. Sapir, eds., *The European Internal Market: Trade and Competition*, Oxford: Oxford University Press, 1989.
- Bhagwati, Jagdish, "Multilateralism at Risk," The Harry Johnson Memorial Lecture delivered in London, July 1990.
- Dornbusch, Rudiger W., "Policy Options for Freer Trade: The Case of Bilateralism," in Robert Z. Lawrence and Charles L. Schultze, eds., *An American Trade Strategy*, Washington: The Brookings Institution, 1990.
- Jacquemin, Alexis, *The New Industrial Organization*, Cambridge: Oxford University Press and MIT Press, 1987.
- _____, "Horizontal Concentration and European Merger Policy," *European Economic Review*, 34, 1990.
- _____, and Sapir, André, "Competition and Imports in the European Market," CEPR Discussion Paper No. 474, London 1990.
- Krugman, Paul, "Increasing Returns and Economic Geography," NBER Working Paper No. 3275, Cambridge, 1990.
- Messerlin, Patrick A., "Antidumping Regulations or Protrust Law? The EC Chemical Cases," mimeo., World Bank, 1989.
- Neven, Damien and Röller, Lars, "European Integration and Trade Flows," CEPR Discussion Paper No. 367, London, 1990.
- Sapir, André, "Does 1992 Come Before or After 1990?," in Ronald Jones and Anne O. Krueger, eds., *The Political Economy of International Trade*, Oxford: Basil Blackwell, 1989.
- Viner, Jacob, *The Customs Union Issue*, New York: Carnegie Endowment for International Peace, 1950.
- Wonnacott, Paul, and Lutz, Mark, "Is There a Case for Free Trade Areas?," in Jeffrey J. Schott, ed., *Free Trade Areas and U.S. Trade Policy*, Washington: Institute for International Economics, 1989.

The Challenges of German Unification for EC Policymaking and Performance

By ROBERT F. OWEN*

Rarely have economists been treated to an economic "experiment" comparable to the one of German reunification. In less than a year, a single political and economic entity is being fused from two economies with fundamentally different underlying principles of economic organization and substantially different levels of economic development. (See Horst Siebert, 1990, and David Begg et al., 1990.) Nonetheless, major regional disparities between western and eastern Germany (what were the Federal Republic of Germany (FRG) and German Democratic Republic (GDR)) will undoubtedly continue for many years. Since the fall of the Berlin Wall on November 9, 1989, both German Monetary Union (GEMU) and political reunification were realized on July 1 and October 3, 1990, respectively. In this process, the advice of economists was often ignored in favor of political imperatives (see Roland Vaubel, 1990).

Concerned with the potential impact of German Economic and Monetary Union (GEMU) on the process of European economic and political integration, many European Community (EC) policymakers have perceived German reunification as adding impetus to the "1992" program. Nonetheless, such initial reactions to GEMU have generally been formulated with only fragmentary information regarding the subsequent economic costs and policy measures, necessitated by the real resource transfer in

order to achieve the German government's stated objective of equalizing living standards between western and eastern Germany. To the extent that GEMU generates positive or negative spillover effects to other EC countries, it presents challenges that can potentially influence the speed and degree of European economic integration.

This paper examines economic implications of GEMU for the European integration process. In so doing, it makes reference to the burgeoning literature on the economic consequences of German reunification. Consideration of statistical information related to the performance of eastern and western Germany since June 1990 permits some evaluation of the validity of certain hypotheses underlying earlier studies of GEMU, as well as an assessment of likely consequences for EC policymaking and performance. Nonetheless, it is contended that the most provocative implications of GEMU result from the specific case it provides for analyzing the costs involved in integrating the fundamentally different centrally planned economies (CPE) of Eastern Europe within a viable European "economic space." In this sense, the German reunification "experiment" offers not only the remarkable opportunity to measure the economic costs associated with the total reorganization and restructuring of a CPE to a free-market economy, but also to assess the contribution of government policy measures in this conversion process. Yet, certain unique characteristics of GEMU limit the validity of lessons from the German experience for understanding the adjustment processes in other Eastern European economies.

I. Conceptual Issues

The macroeconomic consequences of GEMU have been analyzed by Begg et al.

*Associate Professor, European Institute of Public Administration, P.O. Box 1229, 6201 BE Maastricht, The Netherlands; and Researcher, CERCORE, University of Nantes, France. A special thanks is due to Dietrich Hartenstein of the Deutsche Bundesbank for insightful discussions, along with assistance in the identification of German statistical sources. The diligent research assistant of Eugene Canjels, as well as suggestions by Klaus Gretschnann, Jacques Melitz, and Watanabe Shinichi are also gratefully acknowledged.

and by Michael Burda (1990), and their empirical effects simulated in multicountry models by Lewis Alexander and Joseph Gagnon (1990), Paul Masson and Guy Meredith (1990), and Warwick McKibben (1990). A clear consensus, arising from these and other studies, is that GEMU will likely entail some rise in world interest rates and pressures for a real appreciation of the deutsche mark (DM). An associated nominal realignment within the European Monetary System (EMS) has the potential for reducing the credibility of "hard currency" policies that have led in recent years to the convergence of inflation rates in most other EMS countries to those of West Germany, and a narrowing of the range of variations in EMS currencies in relation to the European Currency Unit (ECU). Thus, GEMU has the potential for affecting the prospects for European Monetary Union (EMU)—the final stage of which is targeted for January 1, 1997.

One rationale for a real DM appreciation focuses on the inflationary impact of the additional export demand from East Germany—particularly for investment goods. Given the near-capacity level of many firms in the FRG, real exchange rate appreciation is a mechanism by which German industry would become increasingly specialized in capital goods production, while imports of consumption goods from other countries would increase, but perhaps temporarily. Capital inflows to Germany, induced by favorable investment opportunities and higher real interest rates, could compound this tendency for a DM appreciation, which Begg et al. have characterized as "overwhelming" (p. 65). While such pressures on the DM will depend on the specific fiscal and monetary policies adopted in Germany, a probable scenario is for only limited, additional tax financing and continued restrictive monetary policies by the Bundesbank. Hence, the real resource transfer to eastern Germany, will likely generate a substantial government budget deficit.

A consequent dilemma appears to confront adherents of a fast move toward EMU. Rather than permitting an EMS realignment that could be perceived as weakening

the prospects for attaining the third stage of the Delors plan, other EMS currencies may be allowed to appreciate with the DM. Such a prospect, along with higher real interest rates, suggests a scenario where GEMU entails a slowdown of economic activity elsewhere in the EC. However, in the longer term, other factors, such as a deteriorating trade balance and a lower per capita GNP for an unified Germany (relative to the FRG), could entail a DM depreciation. Thus, the size of the real resource transfer to eastern Germany appears critically related to the potential for short-term, exchange rate overshooting.

The foregoing analysis suggests that the implications of GEMU will be influenced by changes in the external position and commodity composition of Germany's trade, as well as by a likely Germany fiscal deficit. As noted by Daniel Gros and Alfred Steinherr (1990), a high, East German propensity to purchase consumer goods produced by other EC countries dampens pressures for an initial appreciation. A related issue concerns the degree to which increased demand for capital goods and related intermediate components will be skewed towards German producers. Although Begg et al. highlight the dominant positions of West Germany and Japan as net exporters of machine tools and other engineering goods, other European producers, such as Switzerland and Italy, may also be beneficiaries. Excessively high, resource transfer costs due to GEMU could jeopardize German commitment to the additional distributive support for "Southern Europe," provided by the so-called "flanking policies" of the Delors plan for EMU. As analyzed by Paul De Grauwe (1990), the precipitous supply-side shock to the GDR's economy might lead to a "Mezzogiorno syndrome," whereby a permanently disadvantaged region is created in eastern Germany, requiring massive resource transfers for many years. Giorgio Basevi (1990) suggests that such potentially large regional disparities could revitalize the "two speeds approach" to European integration. According to this argument, a divergence could arise in the pace that certain countries would be willing to undertake in-

stitutional changes towards European integration.

In view of severe economic difficulties and uncertainties facing the fledgling democracies in Eastern Europe, GEMU provides a remarkable "experiment" that can offer insights regarding the role and costs of government policies in the transformation process from centrally planned to free markets economies. Major issues concern the viability of producers in eastern Germany, costs of resource transfers, immigration pressures, optimal privatization policies, the responsiveness of private-sector capital flows to investment opportunities, and the efficacy of alternative public investment initiatives. The "shock therapy" of GEMU, wherein eastern German firms must rapidly face national and international competition, may result in a speedy transformation process or generate a Mezzogiorno-type economy. Nonetheless, certain unique aspects of German reunification limit the validity of inferences for other Eastern European countries. The automaticity with which West German legal institutions have been adopted eschews highly conflictual issues related to the assignment of property rights, creating a more favorable investment environment. Other differences include the overriding commitment of the German government to equalize living standards (as represented by the German solidarity fund), the substantial transfer of purchasing power to East German consumers resulting from GMU, along with the monetary discipline of the Bundesbank. Finally, the higher standard of living, stronger productive sector, and lower international indebtedness of the GDR also warrant consideration.

II. Macroeconomic Simulation Results

Econometric simulations of GEMU, including those previously cited, indicate a dependency of estimated results on alternative German policy responses and an hypothetical EMS realignment. Yet, perhaps the most pervasive characteristic of these studies is the relatively small magnitude of the estimated effects. For example, based on simulations of the IMF's global model,

Masson and Meredith found an upward pressure on long-run real interest rates in the FRG of $\frac{3}{4}$ of a percentage point, a modest, $3\frac{1}{2}$ percent real appreciation of the DM relative to the dollar, and a reduction in West Germany's current account surplus equaling 2 percent of GNP. Other effects on the rest of the world were generally small.

Similarly, McKibbin's simulations based on the McKibbin-Sachs dynamic multicountry model, show that if the GDR is initially taken as a "separate economic entity outside the model" (p. 8), an anticipated fiscal expansion leads to an increased German fiscal deficit by 3.3 percent of GDP as of 1991, a 1 percent rise in world interest rates, and a 7.7 percent appreciation of the DM in relation to the dollar. Under the alternative scenario of a unified Germany, McKibbin found a net 8 percent increase to total unemployment after the first year. Nonetheless, such econometric results are based on preunification models that do not explicitly incorporate estimates of either the microeconomic costs associated with restructuring East German industry, or the associated pressures for real resource transfers.

III. Recent Empirical Evidence on the Performance of the GDR¹

Two particularly striking features of GEMU since June 1990 are the surge in exports from western to eastern Germany and the virtual collapse of eastern German industrial output. Specifically, inventory accumulation in anticipation of GMU generated exports from the FRG to the GDR for June equal to 312 percent of the May value. Similarly, overall purchases from West Germany by the GDR, during the 5-month period beginning in June, were 3.57 times those of the corresponding months in 1989. This precipitous jump dwarfs the comparable

¹Unless otherwise indicated, the statistical information cited here is based on published series from Deutsche Bundesbank and the Federal Office of Statistics.

multiple of 1.23 for imports from the GDR, leading over this period to an overall, 8.2 billion DM trade imbalance. Furthermore, accumulated FRG sales of investment goods to the GDR, from June through October 1990, have risen dramatically to a level 3.29 greater than the corresponding 5-month period a year earlier. These constituted 36.3 percent of total exports from the FRG to the GDR over the more recent period. Nonetheless, the increases for food products, raw materials, and other consumption goods exports are also substantial (in corresponding order, 11.50, 1.49, and 35.5 times greater), while their respective shares of the 5-month total are 27.5, 17.2, and 9.3 percent.

The prospect of positive spillover effects for a number of EC countries is also suggested by their improved trade balances with Germany, reducing immediate pressures for a DM appreciation. Specifically, over the 5 months since June, the value of West Germany's imports from the EC rose by a monthly average of 9.5 percent (relative to the same months in 1989), whereas imports of the EC decreased by 3.9 percent. Yet, the substantial variance across EC countries in the average monthly changes in their exports to Germany, as well as differences in the levels and propensities of exports to the German market, suggest that the country-specific effects of trade generation resulting from GEMU need to be examined in more detail. (During the 5 months, the largest monthly increases of exports, again, relative to 1989, were from Denmark, Belgium plus Luxembourg, and Italy, that equaled 16.5, 12.2, and 11.2 percent, respectively. In comparison, the value of exports from Japan rose by 3.1 percent, while those from the United States fell by 8.6 percent.) In fact, there is a quite low correlation coefficient of .17 between the shares that exports to West Germany represent in each EC country's global exports, on the one hand, and the total imports of the FRG, on the other (based on OECD COMTAP trade data for 1987). Hence, it is apparent that any simulated net contribution of GEMU to Germany's aggregate macroeconomic performance can be associated with quite dif-

ferent trade multiplier effects across the different EC economies.

East Germany's industrial output had fallen by September 1990 to a level 51.1 percent below its level in the same month of the previous year. Associated with this dramatic production loss is the more gradual, but steady increase in unemployment. The 589,200 unemployed in November constituted 6.7 percent of the work force in eastern Germany. While the threshold of what will be the lowest level of eastern German production remains to be observed, unemployment in eastern Germany appears to be lagging the more precipitous production downturn. This portends an ominous omen for any realistic assessment of the resource transfers that will be necessary to provide unemployment compensation and subsidies to mitigate continued migration pressures. Indeed, in light of such rapidly rising government expenditures, recent estimates of the German budget deficit for 1991 (cited in the German press) have been as high as 150 billion DM, or between 5 and 6 percent of forecast GNP. This suggests the necessity of reassessing simulations of the global impact of GEMU.

IV. Conclusions

In the course of examining the implications of GEMU for the European integration process, an initial focus on the macroeconomic performance of a united Germany highlighted the potential pressures for a short-term real appreciation of the DM, an increase in real interest rates, a German fiscal deficit, and a foreign capital inflow into Germany. Depending on the commodity composition of the trade generated by GEMU, and the dominant position of West German producers in the capital goods industry, there will also be associated effects on trade flows between Germany and the rest of the world. However, the magnitude of estimated results obtained in recent econometric simulations of GEMU suggest relatively small spillover effects for other countries' performance.

A contention of this paper has been that perhaps the most intriguing aspect of

GEMU is the "experiment" it provides for assessing the adjustment costs involved in converting a centrally planned economy to a market economy. This, in turn, suggests that the validity of using existing models to simulate GEMU ultimately depends on a series of empirical issues related to the postunification performance of East Germany. In this latter regard, initial empirical evidence reveals an unidirectional surge of exports from West to East Germany, the collapse of the supply side of the GDR's economy, an associated build up in the German fiscal deficit, as well as dampened German exports, but substantially higher imports to Germany. The latter appear to be rather unevenly spread across countries, and by no means confined to capital goods. Certain of these effects appear to have been insufficiently characterized by existing empirical work, suggesting the methodological inadequacy of relying on simulation approaches that do not explicitly model the GEMU process. Clearly, the GEMU phenomenon will not only generate intra-German and international trade, but also a substantial research agenda. In this latter regard, one intriguing issue concerns the interrelation between internal and international investment flows for Germany, and the trade patterns identified here.

REFERENCES

- Alexander, Lewis S. and Gagnon, Joseph E., "The Global Economic Implications of German Unification," International Finance Discussion Paper No. 379, Board of Governors of the Federal Research System, April 1990.
- Basevi, Giorgio, "Some Economic Consequences of German Unification: Implications for the European Economic and Monetary Union," manuscript, University of Bologna, September 1990.
- Begg, David et al., "The East, the Deutschmark and EMU," in *Monitoring European Integration: The Impact of Eastern Europe*, London: Centre for Economic Policy Research, October 1990, 31-76.
- Burda, Michael C., "The Consequences of German Economic and Monetary Union," Discussion Paper No. 449, Centre for Economic Policy Research, London, August 1990.
- De Grauwe, Paul, "The Economic Integration of West and East Germany. Two Tales Based on Trade Theory," International Economics Research Paper No. 72, Catholic University of Leuven, October 1990.
- Gros, Daniel and Steinherr, Alfred, "Macroeconomic Management in New Germany: Its Implications for the EMS and EMU," manuscript, Centre for Economic Policy Studies, Brussels, 1990.
- Masson, Paul R. and Meredith, Guy, "Economic Implications of German Unification for the Federal Republic and the Rest of the World," Working Paper No. 90/85, Research Department, International Monetary Fund, September 1990.
- McKibben, Warwick J., "Some Global Macroeconomic Implications of German Unification," Discussion Paper in International Economics, No. 81, Brookings Institution, May 1990.
- Siebert, Horst, "The Economic Integration of Germany—An Update," Kiel Discussion Paper 160a, Kiel Institute of World Economics, September 1990.
- Vaubel, Roland, "German Monetary Reunification: The Economic Consequence of Returning to the Pre-War Parity," manuscript, University of Mannheim, August 1990.

Integration of Eastern Europe into the World Trading System

By HELEN B. JUNZ*

One year after the historic decisions of almost all Eastern European countries¹ to change their economic orientation toward guidance through market signals, the stark realities of that decision have become clear. They are the more apparent as the efforts at economic reform break decisively with the past. Previous such attempts generally sought to find ways of cohabitation for what remained essentially different economic systems, including a way to operate within a multilateral trading framework. Today, these countries no longer seek a way to live in a world with a different economic orientation, but rather to adapt themselves to be part of that world and to regenerate linkages that have been stunted during decades of politically determined, inward-looking economic development. In short, they look to political and economic transformation. Thus, the fundamental question no longer is "why cannot they be more like us?" but, now that they have decided to be like us, "what is the most efficient, therefore least painful, way of doing so?" This leads to the basic question: how deeply rooted are their present regional links, as manifested in the CMEA,² not only in terms of trading relations, but also in terms of employment and production structures, and what rigidities do they bespeak?

This question is virtually impossible to answer with any degree of precision as, in an environment that put a high premium on administrative success, so-called factual information does not necessarily reveal hard facts. Data, thus, are a mixture of fact and fiction. But it is a fact that the CMEA group has traded on the basis of raw material inputs, particularly energy, at below world market prices. This has led to capital stocks and production structures that, in market terms, are wasteful; as capital structures have been maintained well into technological senility, it will not be easy to shift resources from supplying command-economy markets to satisfying market-based demands. Although the breadth of this problem cannot be gauged, indicative numbers show that, for example, Hungary's and the CSFR's output of machinery and engineering products are deeply regionally intertwined, with about 30 percent for use in other CMEA countries. Such numbers show that integrating these economies into the high-tech, highly competitive markets of the West will prove a daunting task.

Why is this so? First, there is the sheer size of intra-CMEA trade (between 40 to 80 percent of members' overall trade). This trade moved largely under long-term contracts and provided a certainty of production runs and, therefore, also of employment. This in turn produced managerial attitudes not well adapted to facing the uncertainties inherent in a competitive situation. Second, production based on these arrangements was not able to earn its way: the trading unit of account, the transferable ruble, was considered by most participants to be highly overvalued against other CMEA currencies; this meant, in many cases, that fulfilling CMEA contracts required varying levels of subsidization from national budgets. Moreover, even though in most countries the price structure floated on low-priced raw material inputs, a large part of

*Special Trade Representative and Director, International Monetary Fund, 58, rue de Moillebeau, 1209 Geneva. The views expressed are those of the author and not necessarily those of the Fund.

¹Bulgaria, the Czech and Slovak Federal Republic (CSFR), Hungary, Poland, Romania and, until mid-1990, the German Democratic Republic.

²Council for Mutual Economic Assistance, the regional economic organization charged with monitoring trade and payments and promoting specialization of production. Its members are: Eastern Europe listed in fn. 1, the USSR, Cuba, Mongolia, Vietnam, and Albania (inactive); Yugoslavia is an associate member; Afghanistan, Angola, Ethiopia, Finland, Iraq, Mexico, Mozambique, Nicaragua, and Yemen have cooperative links; in addition there are numerous observers.

output, because of design and quality problems, could not have obtained even the CMEA price in Western markets.

Integration of these economies into the world trading structure involves the whole complex of structural adjustment policies, including the need to surface, and then to eliminate, the subterranean inflation and unemployment that have cumulated during more than a generation of command-economy management. Thus, the problem is how to do away with multiple pricing, direct and cross subsidization, absorption of resources by noneconomic activity, suppression of inflation and unemployment; in short, stripping away the administrative economic cocoon to reveal the butterfly of comparative advantage. All this, obviously, cannot be achieved overnight, but the risk that opportunity might flutter by is high.

In these circumstances, the main questions are: 1) What is the best institutional environment in which comparative advantage can be revealed, uneconomic activities weeded out, and policy credibility and public support (the two being essentially mirror images of each other) maintained? 2) At what speed should adjustment take place?

In considering these questions, the centrally planned economies in process of transformation often have been viewed as a cohesive group (i.e., a bloc), susceptible to a generalized policy prescription. Clearly, this "bloc mentality" belongs to the past. Each of these countries has different characteristics, for example, with respect to resource endowment and to the institutional and political capacity to support and sustain change; these differences impact the structure and possible speed of economic transformation. Even so, some general points apply.

First, any policy reform package must be comprehensive. This is so because the main elements of a market economy are interrelated; for example, successful price reform is not possible if channels for transmittal of price signals do not exist or are clogged. This means that wage reform, acceptance of financial responsibility by enterprises, and reform of the financial sector must go hand in hand with the move to market prices.

Experience shows that piecemeal reform, such as tried by Hungary and Poland prior to 1989, does not succeed. In fact, a piecemeal approach carries the seeds of policy reversal within itself; such reversal, in turn, progressively undermines effective policy implementation because each such episode erodes policy credibility and, thereby, increases reaction lags by reform-minded entities, while shortening those of vested interests. Thus, clarity of purpose and predictability of policy action are a *sine qua non* for the success of reform strategies.

Second, it is obvious that comprehensive reforms, particularly when distortions are deeply rooted, cannot be implemented overnight. In fact, Adam Smith already saw this:

The case in which it may sometimes be a matter of deliberation, how far, or in what manner, it is proper to restore the free importation of foreign goods, after it has been for some time interrupted, is, when particular manufactures, by means of high duties or prohibitions upon all foreign goods which can come into competition with them, have been so far extended as to employ a great multitude of hands. Humanity may in this case require that the freedom of trade should be restored only by slow gradations, and with a good deal of reserve and circumspection. [1937, p. 435]

While one must agree with Smith that too fast a speed could undermine support for the right policies, it also is true that too slow a driver may cause as many accidents as a speedster. The need to find the realistically safe and sustainable speed for implementation of reforms, however, must not be confused with the tendency to delay politically and socially difficult actions under the name of "gradualism." Thus, the natural uncertainties of the market must not be compounded by uncertainties about policy action; and the pace of implementation must be sufficiently fast to allow an early harvest, but not so fast as to outpace the institutional change without which operational measures cannot have their intended effect.

The need for the reform effort to move forward on a broad front is rooted in the need to generate an investment structure adapted to today's and to future economic facts. But some imperatives of reform are more equal than others. First, if the reform effort is to succeed, economic accountability at all levels, enterprise management, government and individual, is a must, particularly as sanctity of contracts generally has not been a feature of the command economy. De facto accountability has to be accompanied by the development of a judicial system that governs private sector transactions and provides predictability. These are the preconditions for sound longer-term investment and the generation of investable savings. They are the more urgent as the scarcity of available resources, financial, budgetary and real, leaves virtually no room for mistakes.

Third, reform must be accompanied by trade liberalization to help break down domestic monopolies and to gain the efficiencies from the division of labor across borders. In increasing links to the rest of the world, Eastern Europe has looked to close involvement with the three international economic organizations—the IMF, World Bank, and GATT. Most Eastern European countries are GATT members of long standing.³ But their economic systems did not allow them fully to meet the GATT requirements on reciprocity, nondiscrimination, and transparency, and, in the absence of market pricing, caused adoption of special safeguards by other members to guard against possible dumping by state-trading organizations. With adoption of price guidance, these countries should become fully functioning members.

This raises a set of problems, but also opportunities, some of which are closely intertwined with the reform process: how fast can subsidies be dismantled, and what are the consequences of shifting from essen-

tially barter to trade in convertible currencies? In this respect, a central question is whether halfway houses provide positive help toward making these economies internally and externally competitive. As long as a significant part of the existing capital stock remains in use, some sizeable intra-CMEA trade, if only in spare parts, will continue to exist. In addition, the existing infrastructure (transport links, ability to off-load and process raw materials, especially oil) also mandates continuation of CMEA trade. Finally, the shift to market pricing will improve the USSR's terms of trade and associated potential financial strains could impact the intended move away from countertrade. This potential has given rise to proposals for a Central European Payments Union (CEPU) among a varying number of CMEA countries (excluding the USSR). This is seen as easing the road to external convertibility and maintaining what are thought to be natural economic links among neighboring countries. However, as the main regional payments problems would arise vis-à-vis the USSR and the need is to build channels to markets and suppliers outside rather than within a truncated CMEA, the rationale for a CEPU seems not soundly based; in effect, it might lead to permanent occupation of the halfway house. This in no way means to minimize the financing problems that are likely to arise once normal production levels are reestablished and import demand recovers from its currently depressed levels. However, more direct avenues would seem to provide better solutions.

Developments in 1990 bear this out. The fast integration of East Germany into a unified Federal German Republic, which has cut deeply into CMEA flows, efforts by reforming countries to reorient their trade toward Western markets and the sharp falls in domestic output and demand, all have caused intra-CMEA trade to contract. Consequently, it is estimated to have fallen by some 20 percent in volume in 1990. This trend could accelerate with the dissolution of the traditional CMEA and the move, in principle, to trading at world market prices and in convertible currencies as of January

³Bulgaria is in process of accession and the USSR became an observer in early 1990.

1991. In anticipation, a large number of bilateral barter deals are being concluded, both government to government and enterprise to enterprise. However, future trends in intra-CMEA trade are likely to bear little relation to current deals as input needs over the medium term, and legal and managerial decision-making ability of enterprises, are uncertain. Another factor is the past history of elastic interpretation of contracts.

On the whole, anticipation of changes in the regional trade and payments system has aimed to reduce bilateral imbalances, particularly, but not exclusively, vis-à-vis the USSR. This included both exchange rate and administrative measures. The basis for new contracts largely involves cash settlements and supplier credit arrangements, rather than, as earlier anticipated, special clearing arrangements that emphasize regional ties. But, with inter-enterprise relations still embryonic, barter deals may continue to dominate for some time. Even so, the effects of changes in the relative prices of intraregional trade will put further pressure on the pace of adjustment: obsolescence of whole branches of production will surface more clearly and, consequently, also the need to restructure capital stocks; but this, in turn, will complicate the digestibility of adjustment efforts as social pressures, particularly those associated with rising unemployment and inflation, will be exacerbated.

At the same time that old trading relationships disintegrate faster than anticipated, integration into the world trading system remains time consuming. The "early harvest" that followed removal of disincentives to market-oriented trade (particularly in agriculture and consumer goods) probably is beginning to run its course; further export growth, consequently, must rely on restructuring of production, including the ability to respond flexibly to emerging and changing demand stimuli. This is more urgent as the events in the Persian Gulf have shrunk traditional markets and increased world energy prices. Although the latter should help rationalize energy use, it may overload the adjustment circuits.

As recent developments show, the region cannot really look to intraregional trade to cushion the process of integration of its production structures into the world economy. This makes a durable acceleration of expansion of trade relations with the rest of the world yet more pressing. Therefore, it is not surprising that these countries, with Poland, Hungary, and the CSFR in the vanguard, are seeking special access to the markets closest to them. Thus, they are in the process of negotiating association agreements with the EC and look to free-trade negotiations with European Free Trade Area. More broadly, adoption of market pricing and abrogation of state trading, as noted above, provides the basis for full integration into GATT. This involves introduction of nondiscriminatory and largely bound tariffs at economically justifiable levels and, for other GATT members, abolition of their special safeguard clauses. A national tariff structure for Eastern Europe obviously requires that relative prices reflect comparative advantage domestically and externally. This goal is undercut by the external trade restraints that exist in many markets, as these are an impediment to rational resource allocation, especially important during the transition period.

Equally important, the ability to increase exports in the face of large import-intensive investment requirements is crucial if external viability is not to be compromised by excessive levels of external debt. Eastern Europe has a known comparative advantage in agriculture and in certain soft goods and, once technological bars are lowered, will compete in high-tech areas on the basis of a highly educated labor force. These are exactly the trading areas that are rife with structural barriers in the most promising markets in Europe and elsewhere, and where head-on competition is likely to engage producers in both developing and industrial countries. This is an additional element in the effort of Eastern European countries to negotiate special agreements with the EC, which in turn has raised fears on the part of current suppliers both EC-associated (such as Maghreb) and nonasso-

Industry Restructuring in East-Central Europe: The Challenge and the Role for Foreign Investment

By CATHERINE L. MANN*

Two key reforms underpin the moves toward more market-oriented economies in East-Central Europe (ECE): price reforms that allow a freer alignment of relative prices toward international norms; and industry reforms that increase competition, flexibility, and efficiency. These reforms are linked; only competitive firms responding to rational price signals can create an industrial structure able to withstand the rigors of international competition.

I. The Challenge of Changing Industrial Structure

Privatization is the transfer of ownership from the state to the private sector and is the hallmark of the market-oriented reform programs. But a market economy cannot be built from privatized enterprises alone, because the existing industrial structure of the East-Central economies lacks robust midsize enterprises. Midsize firms in industrial market economies transmit price signals and absorb shocks to employment and output. Thus, industry restructuring in ECE requires both privatization and new enterprise creation, particularly of midsize enterprises.

The concentration of ownership in state hands is clear.¹ In the Czech and Slovak Federal Republic (CSFR) as of 1989, 90 percent of industry was state owned and was so monopolized that 60 percent of the inputs to final production were sourced from

single plants. In Hungary, state-owned industrial enterprises accounted for 80 percent of the value of output in 1987, private enterprise less than 2 percent.

A second legacy of the command system is a dearth of midsize enterprises. Although hard to define precisely, the midsize enterprise is between the "atomistic" entrepreneur and the large corporation in terms of production complexity, employee size, and economic power. The midsize enterprise has greater fixed and working capital investment, greater demands for organized labor input (as in a team or on a production line), and requires more management than a small enterprise, but it is too small to wield any power over input or output markets. Such an enterprise could be a subsidiary of a large corporation so long as it remains a separate decision-making unit operating at arm's length.

The midsize enterprise plays several important roles in the market economy. As producers of intermediate inputs and niche-market products, midsize enterprises increase diversity of output and flexibility and efficiency of production. As distributors and wholesalers in internal and external trade, midsize enterprises link the larger firms to each other and to the final consumer. As the competitive core of the economy, they channel and refine the price signals between producers and consumers.

Judging what is an appropriate mix of enterprise sizes for an economy requires knowledge of output markets and production technology of leading industries, and an understanding of the lines of corporate financial and managerial control. Nevertheless, data suggest that East-Central economies lack independent midsize decision-making units in both manufacturing and trade.

In industrial market economies, small and midsize firms are the core of the economy.

*Staff economist, Division of International Finance, Federal Reserve Board of Governors, Washington, D.C. 20551. This paper represents my views solely and should not be interpreted as reflecting those of the Board of Governors of the Federal Reserve System or of other members of the staff.

¹Data are from national statistical yearbooks, interviews with government officials, and translations of newspaper articles and publications of economic institutes.

Considering manufacturing alone, the vast majority of firms (more than 85 percent) employ fewer than 100 people but account for about a third of manufacturing employment and about a quarter of the value of manufacturing output. Midsize firms (between 100 and 500 employees) account for about 10 percent of firms, and a third of both employment and output. Data for the Netherlands (which is similar in population to Hungary and the CSFR) indicate that midsize establishments also account for about a quarter of export sales.

The legacy of the command system makes the enterprise the relevant decision-making unit in ECE. The CSFR has the most centralized manufacturing structure. In 1987, 66 percent of employment was in large enterprises (1000 to 5000 workers). Small enterprises (less than 500 people) accounted for less than 2 percent of employment; data for smaller enterprises are not available.

Hungary has a more diversified industrial structure in large part because industrial cooperatives have been allowed to develop along side the state enterprises. The state industrial enterprise remains the dominant economic unit in manufacturing, employing 85 percent of the work force and accounting for 92 percent of total industrial output in 1988, although there are about equal numbers of industrial state enterprises and industrial cooperatives in manufacturing (somewhat more than a thousand each). Midsize state enterprises (100–500 employees) accounted for 12 percent of all state firms, and only 9 percent of total employment and output. Large state enterprises (greater than 500 employees) accounted for 18 percent of firms, 83 percent of employment, and 40 percent of output. Most industrial cooperatives were midsize; even though their impact on the economy is small, Hungary would have no midsize firms at all without the cooperatives.

Midsize enterprises increase the complexity of linkages throughout an economy, which is why their existence was incompatible with the command system. However, robust growth of midsize enterprises would ameliorate the strains associated with ratio-

nalizing the state-owned sector and would speed price reform.

II. Privatization and Midsize Enterprise Creation

Privatizing the state-owned sector will not create the web-like industrial structure characteristic of market economies, but it is a necessary adjunct. Likewise, legislation supporting privatization does not equate to encouraging new private-sector development. Accordingly, two additional facets of industrial restructuring are the degree to which private investors can restructure existing firms that are inefficiently organized, and the extent to which bona fide new enterprises have equal access to credit, labor, and other inputs, as well as product markets. In a sense, this second facet represents the outcome of the whole reform process.

Foreign investment can play an important role in industry restructuring, particularly through midsize joint ventures. The joint venture may be particularly promising both for foreign investors and for the countries. The foreign investor may prefer the joint venture because he gains immediate access to the target sector without waiting for an existing enterprise to come up for privatization. Moreover, he can bargain for a majority share from the start (less likely in the large privatizations) and avoid many of the nationalism and employment issues associated with large privatizations. Finally, capital required to engage a foothold in a risky foreign market is generally lower with the joint venture.

For the countries, as well, foreign investment in joint ventures may be the best way to obtain the technology, management expertise, and Western contacts they desire. Joint venture activity may point to the most promising external markets. Because of their size, joint ventures create the base for a dynamic midsize sector. Finally, because some of the credit, talent, and inputs are foreign sourced, a joint venture can create a more level playing field for itself, thereby aiding the overall reform process.

Legislation legalizing or broadening forms of new private enterprise and removing lim-

itations on sales, number of employees, and foreign participation was passed in the last 2 years in Poland, Hungary, and the CSFR. All three countries also are privatizing small state-owned establishments and have moved forward on large privatizations.

Structural change is evident in all three countries. In Hungary, between 1987 and mid-1990, the private sector's contribution to overall economic activity increased from 2 to nearly 20 percent. Two thousand new firms with less than 50 employees opened in 1990, more than doubling the number from 1989. In the first 8 months of 1990, economic activity at midsize enterprises greatly outstripped that at small and large enterprises; aggregate industrial production fell 10 percent, but firms employing between 50 and 300 people increased output by 21 percent.

In the CSFR, the number of private enterprises rose 50 percent between June and September 1990 to 340,000. The main areas of growth were in repair, custom work, building, trade, and personal services. Moreover, in advance of the start of large privatizations, some Ministries broke up their enterprises, doubling the number in the industrial branch to 225 and increasing by fourfold the number in the engineering branch to 366.

In Poland, private enterprise has increased even though the macroeconomic downturn has taken its toll. In the first 5 months of 1990 (the worst part of the recession), 10,000 new private enterprises per month were registered. The most successful appear to be in trade. The registered private sector accounted for 8.8 percent of industrial production, but some estimates put its importance at above 20 percent.

One difficulty in tracking the impact of the burgeoning private sector is lack of information. Some private entrepreneurs have reinterpreted the democratic political reforms in the economic realm—they neither register with the taxing authority or submit economic data to the statistical offices.

The difficulty of establishing clear ownership, tensions between equity and efficiency in privatization schemes, the nexus of ownership vs. control issues, and nationalism

have made large privatizations rather contentious. In all three countries, large privatizations are moving forward slowly, but foreign investment appears to have a minority role.

In Poland's large privatization program, shares of the firms are allocated to different uses, including recapitalization of the banks, funding social security, and so on. Foreign investors desiring a large stake may negotiate with the state and workers, but a 40 percent share would appear to be the upper limit. Of the five enterprises offered for sale in December, none allowed majority foreign ownership.

Legislation governing large privatization is still being discussed in the CSFR, but outlines of the program are clear. Although majority foreign ownership is possible, in that between 40 and 80 percent of a firm will be sold to domestic or foreign investors (with the remaining portion distributed to the population via a voucher system), majority foreign ownership is unlikely to be observed in fact.

In an effort to move more quickly and to depoliticize the process, Hungary's State Property Agency solicits bids from private investment banks for the right to privatize a state enterprise. Even so, most of the tenders have had restrictions on the share of foreign ownership. However, Hungary has a midsize enterprise privatization program that allows greater foreign participation.

In addition to transferring ownership from the state, another objective of the large privatization programs is to restructure firms. In the Hungarian and the CSFR large privatization programs, the investor can break up the enterprise before buying it, handing back to the state functions that do not make sense in the context of a restructured firm. Foreign investors are seen as important guides in this restructuring process. But limiting foreign ownership to a minority position may reduce the interest and ability of foreign investors to play this role.

The joint venture, whereby the state and a foreign investor create a new enterprise in an industry sector, is a parallel track for industry restructuring. The foreign firm

leads, usually with majority ownership, while the state facilitates by leveling the playing field, often through tax holidays. Joint ventures are still relatively small in the marketplace, accounting for only a few percent of economic activity, but they have been growing quickly in all three countries. Many are midsize firms in the most profitable sectors.

In Hungary, about 4500 joint ventures were registered as of August 1990, up from only a 100 or so in 1988. Data for 1989 suggest that the joint ventures were more profitable than the state-owned firms in similar industry branches and accounted for just less than 10 percent of total industry profits earned in the economy. By industry branch, 45 percent were in engineering and these were the most profitable. Light industry and chemicals also received foreign capital infusions. Sectors of emerging interest are tourism, software, and financial, consulting, and accounting services. In contrast to the large privatizations, about a third are majority foreign owned, with somewhat less than a quarter fully foreign owned. Early statistics for 1990 are similar although it appears that the average capitalization dropped between 1989 and 1990 from about \$400,000 per firm to about \$225,000 (some of this may represent exchange rate changes). This represents a smaller foreign infusion than the price of a lower share in a large privatization.

In Poland, about 1500 joint ventures have been registered. Of the 30 largest, 4 each are in the food and hotel industries, with other ventures in furniture, agricultural machinery, and environmental engineering. Forty percent have a majority foreign stake, with 25 percent having a 90 to 100 percent foreign ownership share. Compared to Hun-

gary, the asset value is somewhat smaller, with 18 percent having a value of \$60,000 to \$100,000 and 33 percent less than \$50,000.

In the CSFR, the joint venture is the only way to enter the market, since the privatization program has not yet begun. Nine hundred joint ventures were established as of November 1990, with the number doubling in November. Foreign investment appears to be concentrated in expanding sectors: 25 percent of joint ventures were in engineering and 20 percent in ceramics and glass. As of the third quarter, industrial production overall was down by 3.7 percent over the same period in 1989, but the glass, ceramics, and china sector increased production. Twenty percent of all joint ventures are 100 percent foreign owned. On average, the foreign ownership share is 77 percent.

III. Conclusion

Being incompatible with the former system, East-Central Europe has a few midsize enterprises. A robust midsize sector could aid price reforms, absorb some of the output and employment losses associated with rationalization of the large state enterprises, and provide a wider range of consumer goods.

Foreign investment through joint ventures may be the best way to underpin the reform movement. Joint ventures benefit the investor by giving majority control with a smaller investment, and by allowing entry into a market without the burdens of large privatizations. The joint venture benefits the country by developing a midsize sector to channel Western technology, expertise, contracts, and credit to the most dynamic parts of the economy.

ECONOMIC DEVELOPMENTS AND PROSPECTS IN CZECHOSLOVAKIA, YUGOSLAVIA, AND GERMANY

Czechoslovakia: Recent Economic Developments and Prospects

By KAREL DYBA AND JAN SVEJNAR*

Czechoslovakia provides a unique example of a country that became underdeveloped as a result of an externally imposed system. Before World War II, Czechoslovakia was a democracy, with GNP per capita similar to that of Belgium and Austria. Its industries were on the technological edge and its products were known worldwide for their superb workmanship. By 1990, Czechoslovak GNP per capita is estimated by the World Bank at \$3,300, thus being comparable to that of Venezuela, Gabon, and Yugoslavia, but only slightly above one-fifth of that of Austria and Belgium.¹ Most Czechoslovak products are now of mediocre quality and selling at a discount, if at all, in the West.

This remarkable transition occurred over approximately 40 years. During the post-World War II reconstruction, the country was still run as a market economy, although major parts of industry, banking, and insurance were already nationalized. After the 1948 Communist takeover, Soviet-type economic planning was imposed, the remaining private enterprises were nationalized,² and priority was given to heavy industry. Czechoslovak foreign trade was reoriented from world markets toward Soviet bloc countries.

*Czech Minister for Economic Policy and Development, Trida SNB, Praha-Vrsovice, Czechoslovakia, and Professor of Economics, University of Pittsburgh, 4M30 Forbes Quad., Pittsburgh, PA 15260, respectively. The views expressed in this paper are our own and do not reflect the official positions of the Czech government. We thank Vladimir Dlouhy for useful comments.

¹The estimated GNP of Czechoslovakia naturally depends on the methodology used. Some other studies generate higher estimates.

²Private agriculture was collectivized or converted into state farms.

The Czechoslovak government adhered to the Soviet-type planning system faithfully throughout the 1950's. The economic slowdown in the early 1960's led to a series of reform attempts, that culminated during the Prague Spring of 1968 with a short-lived and partial program of price liberalization, separation of economic policy from political decision making, enterprise autonomy, and workers' participation in enterprise management. However, central planning was reimposed after the 1968 invasion and remained virtually intact until the late 1980's.

I. An Historical Overview of Economic Performance

Even if one were to treat the current estimates of Czechoslovak GNP per capita as downward biased, the profound decline of Czechoslovakia's standing relative to advanced market economies is undisputable. As a result, one has to accept the official as well as Western data on Czechoslovak economic growth with caution. The data indicate that the most impressive rate of growth occurred during the First Five-Year Plan (1949-53), when the officially measured net material product (NMP) increased at nearly 10 percent a year. However, this pace proved unsustainable and in the second half of the 1950's, NMP grew at 7 percent annually. The 1961-65 period witnessed virtual stagnation and led to the subsequent reform. Growth resumed at about 7 percent during the 1965-70 reform period and the economy still registered almost 6 percent annual growth during the following 5 years. A major slowdown to 3.6 percent set in between 1975 and 1980 as the first oil shock turned the terms of trade against Czechoslovakia

within the Council for Mutual Economic Assistance (CMEA) and agriculture recorded poor performance.

As can be seen from Table 1, the 1980's witnessed a further slowdown in economic growth. The 1.8 percent growth in the first half of the 1980's reflected the impact of the world recession, rising input prices, and restrictive government policies. However, even in the 1985-89 period, the growth rate was only 2.2 percent a year. With many observers noting that inflation was being increasingly underestimated, the 1980's can arguably be seen as a decade of economic stagnation (see Dyba, 1989).

Other indicators also signaled deterioration in economic performance. The share of net fixed investment in NMP fell from 20 percent in 1975 to a mere 13 percent in the late 1980's, and the share of consumption rose. Export growth slowed down in the 1980's and exports to economically troubled developing countries were increasingly accompanied by trade credits. Within CMEA, Czechoslovakia became a net creditor, especially vis-à-vis the USSR and Poland.

The secular deterioration of economic performance was brought about by a number of factors. The centralization of all economic activity after 1948 initially created phenomenal growth as the command system could rapidly mobilize existing resources. The other engine of growth was the rapid increase in inputs that temporarily sustained respectable growth rate. The shortcomings of the centralized system that gradually became overwhelming were the perverse incentives, limited innovation, inefficient allocation of resources, and rigidities. These latter factors became especially important as input growth could no longer be sustained at high rates, and the quality of marginal inputs declined. The situation was further aggravated by the isolation of Czechoslovakia from world markets and its extreme reorientation on trade within the CMEA. This reorientation contributed to the increasing technological backwardness of Czechoslovak industry and vulnerability to disruptions in the protected CMEA markets.

II. The Start of Economic Transformation, 1988-90

A number of initial reform steps, undertaken in 1988-90, pave the way toward more substantial changes that are to take place in 1991 and thereafter. Even before the November 1989 "Velvet Revolution," the Communist government had reduced the role of central planning somewhat. The November 1989 revolution ushered in a liberally oriented transitional government, and created strong expectations of a radical economic transformation from a centrally planned to a market economy. A number of domestic as well as some external proposals (see Svejnar, 1989) for the strategy of economic transformation appeared, but disagreement also emerged within the government about both the overall direction of the economic transition, and the nature and timing of specific measures. As a result, it was not until May 24, 1990, that a government economic strategy, reflecting the above principles, was officially adopted as a formal resolution.

The June 8-9 parliamentary elections brought about major personnel changes in the federal parliament, with the broadly based Civic Forum and Public Against Violence together winning 170 of the 300 total seats, the Communist party retaining only 47 seats and the Christian Democratic Alliance capturing 40 seats. Less-extensive personnel turnover took place in the executive branch, since many of the ministers of the transitional government belonged to the newly formed coalition of the Civic Forum, Public Against Violence, and the Christian Democrats. The new government in principle adopted the May 24 economic resolution but, apart from the elimination of a negative turnover tax on July 9, 1990, no significant economic measures were adopted in the immediate postelection period.

The main reasons for delaying the reform were the inability to achieve consensus on a specific economic program within the executive branch of the federal government, the desire of the newly elected parliament to play a major role in shaping economic laws

and policies, the need to create a completely new legal framework for economic activity,³ and the onset of difficult negotiations about the relative powers and jurisdictions of the federal and the two national (Czech and Slovak) governments.

On September 1, 1990, the government formally submitted to the parliament a "scenario of economic reform." The document contained the first detailed set of economic and social principles, specific measures, and time parameters. It was also a political document that reflected the major compromises that were speedily concluded within the relatively short period. On the macroeconomic front, the scenario placed top priority on a strict anti-inflationary policy, with all other macroeconomic goals (growth, employment, and balance of payments) being "within reasonable limits" subordinate. Specific measures that were to ensure the success of this strategy in 1990 were a zero growth of money supply and a budget surplus of at least 1–1.5 percent.⁴ Measures proposed for 1991 were more far reaching and included a restrictive monetary policy, a 2–2.5 percent budget surplus, a convertible Crown (Kcs), and a positive real interest rate.

The proposed micro policies stressed the need to induce efficient allocation of resources, introduce new institutions, and minimize the social costs of transition. The main measures identified in this context were 1) a major tax reform emphasizing value-added tax, a personal income tax, and an "enterprise" tax; 2) a budgetary reform stressing independence of units and transparency of budgetary allocations; 3) a process of de-etatization and privatization of property; 4) price liberalization; 5) internal convertibility of the Kcs; 6) reduction and

retraining of redundant labor; 7) legalization of collective bargaining and a stiff tax on wage growth; 8) restructuring of social security and health care systems and their gradual separation from the state budget; and 9) structural (industrial) policy based on Czechoslovakia's comparative advantage.

The proposed measures varied in terms of specificity, consistency, and timing. The tax reform proposal, for instance, listed a detailed set of taxes but, despite the fact that the government was elected for only 2 years, scheduled the implementation of the tax reform over a period of 3 years. The backbone of the tax proposal, the value-added tax, was to be phased in only in 1993.

The proposed privatization package consisted of a rapid auctioning of small properties such as restaurants and workshops, speedy restitution of certain types of properties confiscated by the previous government, and a somewhat slower commercialization and privatization of medium-sized and large firms. The privatization of medium-sized and large firms would rely on a variety of methods, including investment vouchers (to be used by citizens to buy shares of enterprises), preferential sale of shares to employees, sales of shares or entire firms to foreign investors, and leasing of firms. The backbone of the proposed microeconomic program was the liberalization of prices that was to take place for a significant part of commodities on January 1, 1991. In this realm the proposal was both emphatic and cautious, noting the need for various forms of price regulation and the need to link price liberalization with the opening of the economy to imports. The principle of internal convertibility obliged enterprises to sell their foreign exchange to the banks, but promised unrestricted possibilities to buy foreign exchange for international transactions.

The variety of views among the architects of the scenario could be perhaps best gauged by comparing the economic and social sections of the document. In contrast to the subordinate position given to income growth, employment, and social security in the restrictive package of economic mea-

³An alternative would have been to adopt temporarily a modified set of Western (for example, German or EEC) laws. However, in view of the voluminous nature of Western legal statutes and the paucity of skilled translators, it turned out to be simpler to create a new set of Czechoslovak laws.

⁴These goals were in fact pursued from the start of 1990.

asures, the social program stressed the social and ecological orientation of the upcoming market economy and placed priority on social justice, employment, workers' incomes, and social security. It declared significant decline of real incomes to be unacceptable and called for an *ex ante* social agreement on the acceptable limit. This aspect was largely brought in line with the more austere economic part of the reform package by the end of 1990.

The parliament approved the scenario, but, from the standpoint of implementation, the striking feature was the large number of laws and decrees that were to be drafted and passed before the major parts of the reform would be launched on January 1, 1991. This indeed proved to be a major burden and the resulting fatigue was increasingly visible. The introduction of some widely expected measures (for example, the privatization of small enterprises) has consequently been delayed.

III. Economic Developments in 1989 and 1990

As shown in Table 1, the 1989 slowdown in NMP growth turned into a decline in 1990. Employment also decreased, but in a less pronounced way. In many cases the labor force reduction has taken the form of retirements and termination of guest workers.

Despite the commitment to a restrictive monetary policy, various measures pointed to a relatively fast expansion of money supply in the first half of 1990, followed by a decline in the last quarter. Consumer prices rose significantly, reflecting both the increased money supply and the rising velocity of circulation of money. Nominal wages rose slower than inflation, but real consumption continued to increase in the first half of the year. Personal savings have registered a significant shift into foreign exchange accounts. The runs on stores that took place in the last quarter of 1990 suggest that consumers have intensified attempts to reduce their savings and money balances in the expectation of higher inflation.

TABLE 1—CZECHOSLOVAKIA:
BASIC ECONOMIC INDICATORS^a

	Annual Average			
	1980-84	1985-89	1989	1990
1) NMP %Δ	1.8	2.2	0.7	-3.5
2) Employment %Δ	0.6	0.8	0.4	-2.5
3) Official CPI %Δ	1.9	0.6	1.4	17.0
4) CPI (PlanEcon est.) %Δ	3.6	2.6	3.4	
5) Nominal Earnings %Δ	2.2	2.8	2.5	3.6
6) Personal Consumption %Δ	0.8	2.7	1.8	-1.3
7) Social Consumption %Δ	4.1	5.3	7.3	3.8
8) Net Fixed Investment %Δ	-6.8	2.4	3.1	-12.3
9) Money Supply				
M1 %Δ	5.8	4.7	2.7	-0.2
M2 %Δ	4.4	3.3	-0.8	-3.8
M3 %Δ				4.4
10) Exchange Rate				
Kcs/\$Commercial	14.40	15.05	15.05	28.00
Kcs/\$Tourist	—	—	—	28.00
Kcs/\$Parallel Mkt.	29.10	33.70	47.40	—
Kcs/\$Auction			121.24	39.40
Kcs/Ruble Commercial	12.24	10.80	10.00	9.00
11) External Debt ^b \$mill.	5.766	6.406	7.915	8.100

Source: Czechoslovak Government and PlanEcon Inc.

^aUnless indicated otherwise, all values are in constant prices in Kcs;

^bIn convertible currencies.

The point of major uncertainty is the enterprise sector. The fall in production, profits, and trade with CMEA countries all reflect the vulnerability of enterprises to the disintegration of the CMEA market. Indeed, after falling 9 percent in 1989, real exports to the Soviet Union declined additional 15 percent in the first half of 1990.⁵ The concomitant drop in the Soviet deliveries of oil and raw materials resulted in a

⁵By the end of 1990, many enterprises were still unable to conclude the usual annual trade contracts with their Soviet partners. Moreover, those that did were unsure whether the contracts would be honored.

14 percent decline in Czechoslovak imports from the USSR in the first half of 1990. Exports to developed Western countries rose by about 7 percent in real terms, but imports from this area jumped by more than 21 percent from January to June, 1990.

A major problem is enterprise insolvency—it has been rapidly rising. This phenomenon is particularly severe because many government subsidies to enterprises were in the past extended in the form of revolving credits. The current pursuit of tight monetary and fiscal policy has resulted in a major increase in interest rates and requests for rapid repayments of enterprise loans by banks. Enterprises have been caught off guard, as the bank interest rate on short-term loans has recently exceeded 20 percent, and they are increasingly relying on inter-enterprise credit.

Finally, the October devaluation of the Kcs/\$ commercial exchange rate from 16 to 24 has created considerable difficulties for nonexporting firms. Similarly, the 10 percent revaluation of the Kcs vis-à-vis the ruble in the first quarter of 1990 created financial difficulties for enterprises with a heavy export orientation toward the Soviet market. Most enterprises expect to continue operating on this market and have not made significant steps toward diversification or restructuring of operations.

IV. Prospects

At the end of 1990, the official government forecast for 1991 predicted a 5 percent decline in real GNP, a 30 percent rise in the consumer price index, unemployment of 5–7 percent, and a \$2.5 billion increase in foreign debt. The extent to which this forecast will materialize naturally depends on a number of factors.

The short-term economic situation will be significantly affected by key external developments such as the political and economic situation in the Soviet Union and the world price of oil. If extreme, these developments could temporarily swamp the impact of internal systemic and policy changes. In particular, a major disintegration of the Soviet

economic activity, resulting in a decline of paid orders for Czechoslovak goods or of deliveries of Soviet oil, would result in a major shock to the Czechoslovak economy. Of course, certain shock is expected in any case as on January 1, 1991, Czechoslovakia starts conducting its trade with the Soviet Union in convertible currencies and at world prices. This shift results in a significant deterioration in the terms of trade for Czechoslovakia.

On the internal side, the major danger lies in further delays of the reform and in the introduction of an inconsistent package of measures. As the major reform steps are approaching, there appears to be rising discontent, a decrease in credibility and increasing resistance to a significant (albeit temporary) fall in living standards. Further delays would probably solidify the opposition to undertaking a major and rapid economic transformation. Similarly, if the government were to abandon the overall program and enact only a set of partial and inconsistent measures, the resulting disequilibria could engender significant opposition to further reforms. Examples of possible mishaps include the loss of macroeconomic stability, a major rise in inflation in the presence of rigid wage controls, an inability to spur the development of small and medium-sized (private sector) enterprises as major layoffs take place in state-owned firms, undertaking only partial price liberalization, and enforcing the foreign exchange surrender without giving enterprises the possibility to obtain hard currency on demand. These and other partial measures have backfired in other countries, and there is no reason to expect Czechoslovakia to fare differently.

There is also considerable room for optimism. Czechoslovakia is one of the few reforming economies that enjoys relative financial stability, a low level of foreign debt, a solid human capital base, and low wages. The economic program has been relatively carefully prepared and could yield rapid results. Recent investments by foreign companies (for example, Volkswagen and Glaverbell) into Czechoslovak enterprises

Economic Development in Yugoslavia in 1990 and Prospects for the Future

By JANEZ PRASNIKAR AND ZIVKO PREGL*

In this paper, we evaluate the performance of the Yugoslav economy within the framework of reforms currently taking place in other socialist countries. We first present the stream of events leading up to 1990, followed by the development of events in the year 1990, and then discuss the possible scenarios of further development, considering the present situation in Yugoslavia.

I. Economic Performance During the 1980's

A. *The Zig-Zag Economic Policy*

Yugoslavia reached high rates of economic growth during the 1970's.¹ This was achieved primarily on account of large external debt. The growth of the external debt at the end of the 1970's therefore demanded decisive steps to correct economic policy.

During the first period that was the phase of curtailing of domestic demand and stimulating exports (from 1982 to 1984: III) (see V. Bole and M. Gaspari, 1991), economic policy concentrated on reducing domestic demand and stimulating sales to markets abroad (i.e., export promotion, exchange rate adjustment). This was attempted by using a host of restrictive measures including wage and price controls, limits on public energy usage, and controls on investment. This policy had a number of significant consequences: there was no significant increase in the inflation rate, the trade and current

accounts were improved, and domestic demand declined. Nevertheless, the policy had a negligible effect on relative prices since it was based on restrictive measures (a stop-and-go policy).

The subsequent period was the period of policy slackness (1985–86), that nullified the results attained during the first period. This policy attempted to stimulate economic output in the hope that an increase in productivity would bring about economic stability. In order to achieve this goal, an expansionary monetary policy was put in place, interest rates were pegged (with the real interest rate, in effect, negative), and the exchange rate was fixed to an artificially high level that stimulated imports. Predictably, the result of this policy was a rise in the inflation rate.

The third period was the period of full indexation (1987–89) when a new accounting technique was put in place. This encouraged economic entities to adopt a nominal indexation of economic categories. The change in input prices affected only the inflation rate, and had no effect on relative prices.

A comparison of the three periods reveals the vacillating (zig-zag) decisions made by economic policymakers. Shock therapy (active policy measures concerning the exchange rate and interest rates, freezing of wages, and a restrictive monetary policy) is usually followed by a period of increased demand that essentially nullifies the initial results, and will again be followed by a period of economic shock.

This comparison shows that the long-term stagflation in the Yugoslav economy was not a result of misguided economic policies, but of the consequences these policies had on economic performance (see Table 1). The source of this failure is the present economic system and its structural disproportions that are inherent in all socialist coun-

*University of Ljubljana, Kardeljeva ploscad 17, 61000 Ljubljana, Yugoslavia, and Deputy Prime Minister of Yugoslavia, respectively. We thank Marko Jaklic for research assistance.

¹The average rate of growth in GNP was 7.0 percent for 1963–73 and 5.6 percent for 1974–80. The rates were similar to those achieved by similar countries in the world (middle income oil importers). See Jozse Mencinger (1989).

TABLE 1—SELECTED ECONOMIC INDICATORS
FOR 1985–89

	1985	1986	1987	1988	1989
1) Gross Soc. Prod.	0.5	3.6	-1.0	-1.6	-0.8
2) Unemployment ^a	13.7	13.9	13.6	14.1	14.8
3) Retail Prices	77	91	118	199	1256
4) Trade Balance ^b	-1.6	-2.0	1.1	-0.6	-1.3
5) Curr. Acc. Balance ^b	0.3	0.3	1.1	2.2	2.0
6) Total Debt ^b	18.2	19.2	20.5	18.7	18.5
7) Curr. Public Sec. Deficit ^c	2.8	2.8	10.5	5.8	3.9

Sources: *Statistical Yearbook* (1989), *Index* (1990/9).

^aPercentage of the eligible population.

^bIn billions of \$US.

^cPercentage of Gross Social Product.

tries.² Another contributing factor may be the pronounced differences in the development of different regions within Yugoslavia.³ Many arguments have been presented in support of the thesis that it is not possible to achieve long-term stability in the Yugoslav economy merely with macroeconomic policy measures. For macroeconomic policy to be effective, it is necessary to fundamentally change the present economic system.

B. From High Inflation to Hyperinflation in 1989

The realization that fundamental changes were necessary in the economic system resulted in the implementation of the "Three Nominal Anchors" program, introduced in 1988.⁴ The new government was established

²For example, Ivan Ribnikar (1989) states that social ownership is the main cause of the continuing economic problems.

³The difference in GNP between two of the country's regions, Slovenia and Kosovo, increased from 5:1 in 1955 to 7:1 in 1983 (see Mencinger).

⁴The essentials of the program, jointly established with the Yugoslav government and the IMF, were to restrict the growth in nominal wages, loans, and public spending. However, within 1½ months, the restrictions on loans were already rescinded, and by the end of 1988, restrictions on wages followed a similar route.

in March 1989, and decided first to implement institutions based on principles of market economies. They later attempted to stabilize the economy by implementation of shock therapy. The government directed the economy by liberalizing prices, wages, imports, and personal foreign currency accounts, which together would bring about equilibrium of relative prices. In addition, it deregulated the corporate sector, enabling economic entities to operate more autonomously. Finally, the government embarked on a policy of diminishing the budgetary burden on the economy.

Prior to the enactment of any of the stabilization policies, inflation in Yugoslavia had continued on its own rapid course. As mentioned, the new cycle of price increases caused a phenomenal adaptation in the accounting system in 1987. This necessitated a full indexation of all nominal monetary figures. Unfortunately, this practice was not suitable for effective enactment of macroeconomic policy (see Bole and Gaspari). In 1987–88, escalating prices particularly encouraged budget deficits, financed through the help of monetary authorities. The deregulation of wages in 1989 and the policy of accommodating exchange rates resulted in an uncontrollable dance of adjusting prices, wages, and exchange rates, and, of course, heightening inflationary expectations. Without formal institutions to ensure the proper collection of taxes and the resulting problems in financing the budget (the Tanzi effect), and because of the losses of the central and commercial banks, a large number of economic units sought to ensure their continuing income through the help of inflation. Demand for currency also increased during 1988–89, because the demand by enterprises and households increased in real terms. This increased demand for real balances made the printing of money uncontrollable. By the end of 1989, inflation in the Yugoslav economy turned into full-fledged hyperinflation. However, the deregulation of the economy contributed to growth of exports and to increased foreign currency deposits in Yugoslav banks, which directly increased foreign reserves.

II. The Economic Reform of 1990

A. The Program of Economic Reform

Observing that the prior deregulation of the economy was already yielding some desirable results, and that the situation at the end of 1989 required strong measures, in the middle of December 1989, the government prepared "The Program of Economic Reform and its Execution of 1990." The proposed program to curtail the inflation rate (the program of "Four Nominal Anchors") rested on the foundation that the main reason for the inflation was the excessive amount of money in circulation, a result of the central bank's role in the budgetary process. By linking the exchange rate of the dinar (after the denomination of the dinar to 1:10,000) to the value of the deutsche mark (7 dinars for 1 DM) and the resulting internal convertibility of the dinar, the program provided an effective disincentive to the excessive printing of money. Wages were temporarily frozen while most prices were still allowed to adjust freely (around 85 percent of them) and imports were liberalized (about 90 percent of all goods). Simultaneously, the value of the dinar indirectly appreciated in the trade with Eastern European countries, since the central bank stopped its practice of financing the annual surplus by printing money.

In addition, the following measures were also proposed. 1) The National Bank of Yugoslavia should restrict its practice of lending to certain sectors (the federal budget, agriculture, exports) and instead become an independent professional institution whose primary purpose is the control of monetary aggregates. Losses resulting from exchange rate inequities should be covered solely by fiscal (budget) expenditures. 2) Fiscal policy should play a more active role in the stabilization of the economy and should assume the responsibilities that would be abandoned by the central bank under the first measure. 3) The country should continue to reschedule its foreign debt and continue to negotiate the terms of payment with its creditors.

TABLE 2—FLUCTUATION OF IMPORTANT AGGREGATES FOR 1990

	(1)	(2)	(3)	(4)
Jan.	17.3	-0.96	-27.5	6974.6
Feb.	8.4	-0.96	30.4	7400.5
Mar.	2.6	-1.36	1.7	7746.3
Apr.	-0.2	-2.27	20.5	8317.2
May	0.4	-2.72	-15.1	8549.6
June	0.2	-2.17	22.9	8733.7
July	4.9	-0.99	13.6	9257.8
Aug.	1.9	-0.54	3.2	10110.5
Sept.	10.9	-0.38	2.7	9996.0
Oct.	8.1			9467.2
Nov.				
Dec.				

Sources: *Index* (1990/10); Privredna kretanja, Economic Institute of the Department of Law, Ljubljana; and Internal Sources in the National Bank of Yugoslavia.

Note: Col (1) = Monthly Inflation Rate; Col. (2) = Monthly Growth in GNP; Col. (3) = Monthly Wage Increase; Col. (4) = Foreign Reserves in Millions of \$US.

B. Implementation of the Economic Reform of 1990

The program to halt inflation began to take effect by the end of December 1989. The results are evident from Table 2. The data shows a radical reduction in the rate of inflation in the first half of 1990. The monthly inflation rate (64.3 percent in December 1989) was reduced to virtually zero in April and May. It was in these months that full internal convertibility of the dinar was established. Foreign currency reserves actually increased by \$3 billion during the first half of 1990 because of an influx of foreign currency (caused by workers' remittances and the time needed for enterprises to bring foreign currency into the country was shortened) and because citizens traded in their foreign currency for the new, more stable, dinar.

However, during the first 6 months of the year, it was evident that certain goals were not obtained. 1) Wages increased at a monthly rate of 5.5 percent. Economic policy did not succeed in stopping wage increases as predicted. 2) Public spending was not decreased, since resource allocation was still adapted to the conditions of hyperin-

flation and, in large measure, depended on changes in wages. 3) Reform of the banking system did not proceed as anticipated due to problems with the revision of bank balance sheets and difficulties in parliamentary procedures. Interest rates for short-term loans (40–50 percent) constitute a substantial pressure on prices. 4) Industrial production decreased by 10 percent during the first half of 1990 (compared with the first half of 1989). 5) Inflation was still relatively high during the first 3 months of 1990 and increased the real exchange rate of the dinar, thus jeopardizing the export-oriented part of the Yugoslav economy.

These negative trends continued into the second half of 1990, despite the introduction of the second phase of reform.⁵ Inflation in the economy is once again approaching dangerous levels. One reason for this is increasing wages and public spending. Income in the public sector rose by 44 percent in September 1990 in real terms, compared to the same period in 1989. Output is generally decreasing at a slower rate than in the first 6 months of the year, but unemployment is accelerating. The economic decline is also evident in international trade and in the balance of trade. The trade deficit for the first 9 months of 1990 was already \$2.415 billion. By the end of August, foreign currency reserves had grown to \$10.1 billion. However, depositors withdrew 10 percent of their foreign currency savings in October. Depositors had obviously lowered their expectations about the prospects of success for the current reforms.

In large measure, these negative trends were attributable to the reforms themselves. Monetary authorities had abandoned their restrictive monetary policies during the summer months. Financial institutions again actively provided insecure loans, short-term loans to cash-starved enterprises, and advances to bureaus of exchange during July and August. This return to undisciplined activity was inadvertently made possible

through policy measures enacted by the government. In its efforts to stop declining output, the government had eliminated some restrictions on monetary policy that brought about this new wave of undisciplined financial activity. Another contributing factor was increases in wages for government employees that resulted in wage increases throughout the economy. While the increases for government employees were granted in compensation for a virtual freeze in government wages that had been in effect the first 6 months, the public did not look upon these wage increases favorably.

These economic policy measures (increase in selective loans to agriculture, increases in wages at federal institutions, balancing of the budget) along with a growth in the level of wages, increased overall public spending, and the return to financial irresponsibility significantly damaged a recovering economy. This damage, when coupled with the inevitable return to restrictive monetary policy, portends an even larger recession. In addition, a number of external factors also hindered the reform process, including the Persian Gulf crisis (\$1.3 billion in 1990), a drought in a main agricultural area of the country (\$1.6 billion), and heightened centrifugal forces resulting from neglect of federal laws and the practice by some autonomous provinces in establishing their own policies on prices, wages, and fiscal activities.

III. Possible Scenarios for the Future

Any predictions about the future economic developments in Yugoslavia are dependent on political and nationalist questions. Rather than speculate on the outcome of the tense nationalist situation, we can only present the options that seem most likely to us at the present.

The first scenario is that the government believes that democratic processes will continue to spread across all of Yugoslavia. It also believes a democratic environment will permit a successful reimplementation of the reform program. The second scenario is that of the federal presidency, which has forced

⁵The new legislation more courageously promotes ownership and restructuring of the banking system, and discourages monopolies.

the republics to make a decision on whether they would prefer a confederate Yugoslavia or a federation similar to the present state. Finally, some republics are working toward independence, in the belief that they can more successfully enact a stabilization program on their own than in the union. Because Yugoslavia is a country in transition politically as well as economically, all of these options are legitimate and possible.

Although it is impossible to predict which option will eventually be realized, it is clear that any of these options could lead Yugoslavia on a path toward political pluralism, a market-oriented economy, and thereby create an environment in which a consistent economic policy can be successful. Of course, policies that do not result in

these outcomes will not be acceptable to most Yugoslavs and to the international community.

REFERENCES

- Bole, Veljko and Gaspari, Mitja**, "The Yugoslav Path to High Inflation," forthcoming 1991.
- Mencinger, Joze**, "The Yugoslav Economy: Systematic Changes, 1945-1986," in *The Carl Beck Papers in Russian and East European Studies*, Pittsburgh: University of Pittsburgh Press, 1989.
- Ribnikar, Ivan**, "Social Ownership and Financial Systems," in *Introduction to the Financial Economy*, Ljubljana: Pegaz, 1989.

The Economic Integration of Post-Wall Germany

By IRWIN L. COLLIER, JR. AND HORST SIEBERT*

It would be hard to devise a better controlled experiment for comparing different economic systems than the experience provided by East Germany and West Germany: two nations that formerly were one, occupied by people of the same background, the same culture, and the same genetic inheritance, torn apart by the accident of war. On one side of the Berlin Wall is a relatively free economic system; on the other side, a collectivist society.

[Milton Friedman, in W. Breit and R. W. Spencer, 1986, p. 89]

Modern experimental economics cannot duplicate experiments of the scale or the duration of postwar Germany. Now post-wall Germany offers a fascinating follow-up experiment to test the economic integration of an ex-centrally planned economy into a larger, established market system. In this paper we examine the experiment that began with the monetary, economic, and social union (MESU) of East and West Germany in July 1990.

The MESU treaty was signed on May 18, 1990. Two bold strokes resolved the issues of the institutional framework for economic integration and macroeconomic stability. The economic order of the FRG was transplanted to the GDR and East German fiscal and monetary sovereignty was transferred to West Germany in advance of the political act of German unification.

Two processes will determine the size of the post-wall integration dividend. The first process is the privatization of the East German economy—the assignment of private rights to existing East German assets. The second process involves real economic adjustments, including adjustments by existing

firms and the creation of a broad base of new small and midsize businesses. The gains from integration arise from merging a large open economy, relatively well-endowed in capital and technology, with a smaller, semi-autarkic economy, relatively well-endowed in labor and land. While the long-run prospects for Germany look excellent, problems encountered in the privatization and the adjustment processes make it unlikely that the integration of the post-wall German economies will be quick and easy.

I. Initial Conditions and Long-Run Prospects

According to Western estimates of relative real consumption based upon detailed consumer price comparisons and expenditure data for both East and West Germany, real consumption expenditure by East German working families was only about half the West German level (Collier, 1989). Because of higher labor force participation rates, longer work weeks, and shorter vacations in the GDR, average East German living standards were much less than half the West German level during the last half of the 1980's. The ecological gap between East and West Germany was at least as wide.

Aggregate investment in East Germany dropped sharply during the first half of the 1980's, and one could sense a steady deterioration of East Germany's infrastructure and housing stock over the past decade. There were clear indications that planned investment was inadequate to sustained planned growth of output (Collier, 1987).

Looking beyond the crumbling buildings, obsolete equipment, and desolate state of the environment, the long-run prospects for East Germany are bright. Within 4 or 5 years, we expect to see major progress in living standards and productivity as well as effort at ecological recovery. The contributing factors for this rapid growth scenario

*Department of Economics, University of Houston, Houston TX 77204, and Kiel Institute of World Economics, Duesternbrooker Weg 120, P. O. Box 4309, D-2300 Kiel 1, FRG, respectively.

may be identified as 1) the economic integration effect, 2) the capital accumulation effect, and 3) the impact of the market economy on economic behavior.

Economic integration of East and West Germany will exploit differences in factor endowments. West Germany contributes abundant capital and technology whereas East Germany brings a dowry of well-trained (though not necessarily cheap) labor and land. Thus, there will be gains from increased specialization as the new states open themselves more than ever before to exploit trading opportunities with West Germany and the rest of the world. The inability of East German industrial products to compete successfully in international markets in the past was largely due to insufficient specialization in production that limited economies of scale and to a lack of innovation. Furthermore, almost all East German producers were at the mercy of single domestic or CMEA suppliers for critical materials and components. Finally, the regional structure of East German foreign trade, especially its extremely large share of trade with CMEA partners, reflected political constraints rather than economic logic.

The capital accumulation effect is the positive association of high rates of growth with high rates of capital accumulation for economies that start from relatively low endowments of capital per capita. The capital stock East Germany brought into united Germany is largely economically and ecologically obsolete. Almost 55 percent of industrial equipment was installed over one decade ago, compared to about 30 percent in West Germany. This is somewhat comparable to the situation of the West German economy in 1948 when high rates of capital accumulation were necessary to rebuild a war-damaged capital stock, and to equip a labor force that had grown by approximately twelve million eastern refugees (Siebert, 1990).

The final reason for the favorable long-run prospects of the new states of the FRG was demonstrated in the controlled experiment mentioned at the beginning of this paper: economic system matters a great deal for economic performance. Following the eco-

nomic and monetary reform of 1948, industrial production increased in the Bizone (British and American zones) at an annual rate of 50 percent within 5 months. From the second half of 1949 through the end of 1951, real GNP grew at 17 percent annually in West Germany. Were history to repeat itself, the new federal states could make up for much lost time by 1995.

One must resist the temptation to overstate common patterns between West Germany in 1948 and East Germany in 1990. The most important difference between the two starting points is that East German firms entered German MESU as inefficient producers. The adjustment to the market economy in the new federal states entails a J-curve with a substantial drop in production and elimination of unnecessary jobs. In November, there were 589,000 unemployed—an unemployment rate of only 6.7 percent. But a far more accurate reflection of idle labor resources and a forecast of 1991 unemployment in East Germany is found in the 1.77 million workers on short time.

The wage-setting process is also different now from 1948. The wage level in East Germany at the start of economic integration is completely out of line with productivity, and wages are likely to grow even faster than productivity. Without an exchange rate to depreciate that would increase the competitiveness of East German firms, the wage rate should have taken over the function of giving East Germany a temporary comparative advantage in labor-intensive industries. However, as workers continue to move to higher paying jobs in the West and unions in East Germany (with the support of unions in the West) continue to push for wage increases to close the East/West wage gap, cheap labor will be only a fleeting advantage that the new states possess for attracting investment.

The social union, that went into effect simultaneously with the monetary and economic union in July extended a social welfare system that has evolved to fit a relatively rich, well-proportioned economy to cover a much poorer economy in the middle of significant real adjustments in the transition from central planning to a market sys-

tem. Direct investment in existing firms that require large cuts in the workforce will be discouraged because the dismissal constraints represent significant entry costs.

II. Problems of Transition—Reorganization and Privatization

The core of the adjustment process is the restructuring of East German industry. In the restructuring of existing firms, three different aspects need to be distinguished: legal independence of the constituent enterprises in the large industrial trusts (Kombinate); economic efficiency of the individual enterprises; and ownership.

Most of East Germany's production (1988) was organized into 316 Kombinate, of which 221 were vertically and horizontally integrated industrial Kombinate. The Kombinate were protected from domestic competitors by governmental market delimitation, while an extreme policy of planned specialization within CMEA protected the Kombinate from foreign competitors as well. The first stage of reorganization was completed when all the constituent firms of the Kombinate (approximately 8,000) were declared legally independent economic units by law.

Now subject to formidable competitive pressures, East German firms must increase their efficiency by restructuring their input and product mixes: the production of inputs in-house that are cheaper to buy from other firms will be discontinued; departments dedicated to prolonging the life of economically obsolete capital will be shut down; product lines motivated by political considerations will be abandoned; and the provision of social services such as kindergartens will be eliminated. Cutting back the workforce is clearly the most sensitive issue facing managers. Through the end of June 1991, managers have the option of keeping people on the payroll as short-time workers with a large part of the difference in earnings paid by the government. This is the reason why industrial production has fallen by half without a corresponding jump in unemployment.

The organizational restructuring of East German firms is closely linked to the issue

of privatization. The rights to publicly owned firms were transferred to a new government trustee agency, the Treuhandanstalt. The primary task of the Treuhand is to privatize East German firms. The Treuhand has a secondary task that is to restructure industry, and if the net value of assets (i.e., privatization receipts less costs of restructuring) is positive, to transfer the net receipts from privatization to either the governments or the people of the new federal states. Thus the legislative intent was that the Treuhand should only exist for the time necessary to accomplish these tasks. However, there is a genuine risk that the Treuhand might evolve into a super "machinery" of sectoral policy exposed to strong political pressures.

Should the Treuhand restructure firms before offering them to private buyers or sell each firm "as is" and let new owners do the restructuring? This has been the subject of considerable public debate in Germany, and apparently within the Treuhand as well. The restructure-first approach has been rightly criticized on Hayekian grounds that the Treuhand has insufficient information for selecting winners from losers for the allocation of scarce resources generated by the Treuhand's borrowing, or from revenue generated in earlier privatizations. The Treuhand has also met with severe criticism for not proceeding faster with privatization. In 1990, Treuhand sales amounted to only about 2.5 billion DM.

The reorganization of existing firms is just one aspect of the restructuring of the East German economy. An important task of restructuring is to encourage the creation of new and small firms. New firms are needed to absorb the employees laid off as inefficient firms cut employment or shut down altogether. The distribution of firms in manufacturing by number of employees in the GDR in 1987 was characterized by a relatively small share of firms with fewer than 100 employees (4.4 vs. 35.9 percent in the FRG).

Over 226,00 new businesses have been established in the first 10 months of 1990. About half of the new businesses are in retail trade and restaurants. Strengthening this growing *Mittelstand* will be the former owners (or their heirs) of small reprivatized

enterprises. About one-fourth of the 12,000 small firms previously nationalized have been transferred back to the original owners.

German economic integration constitutes a Schumpeterian "event" or a "new frontier" in the sense of Alvin Hansen. Vast quantities of investment will be needed to get restructured firms started and to create new jobs. While a strong boom in West German investment is underway, nothing comparable can be seen in the East as of yet.

The German MESU treaty was drafted with two primary goals in mind. First, it was hoped that the introduction of the DM would check the out-migration of East German labor by changing expectations of future living standards in the East. After dropping to a net out-migration of 12,000 per month in May and June, the flow increases to 30,000 per month during July through September. Second, by transplanting West German economic law, the treaty gave investors a legal environment for investment identical to that in West Germany, at least in principle. The May 18 treaty was supposed to signal the start of the great capital race into East Germany. However, with the notable exception of West German financial institutions and retailers, both West German and foreign investors have stayed pretty close to their starting blocks.¹ Their caution is not ill-founded.

Uncertainty in titles to land. Returning land titles to original owners sounds like a simple and obvious way to reestablish property rights in land. However, the passage of time has complicated matters considerably, and public records of land ownership turn out to have been badly neglected in East Germany. Establishing title to any particular property can be quite difficult, especially at a time when millions of people want to check these public records simultaneously! There is a pressing need for some form of title insurance in East Germany now.

¹The five leading German economics research institutes estimated in October that 5 billion DM of direct investment would flow into the East German economy during 1990 and that another 15 billion DM would follow in 1991.

The infrastructure bottleneck. The run-down public capital stock of roads, railways, airports, and telecommunications is wholly inadequate for the rapid development of the new federal states. There are two issues in widening the infrastructure bottleneck: how fast can the East German infrastructure be expanded and how will the expansion be financed? Lags in supplying public infrastructure investment can be both long and variable—Germany is no exception. Considering the time factor alone, there is a strong argument for privately financing as much infrastructure investment as possible. The argument is considerably strengthened by the prospect of huge future government budget deficits.

Space for new businesses. This is an infrastructural bottleneck that has major consequences for the rate of new job creation. New firms must be able to acquire office space, warehouse space, and locations for setting up production facilities. New industrial and office parks as well as shopping centers are sorely needed.

III. Macroeconomic Aspects of German Unification

The united German economy is divided into an economic boom in the West and an enormous contraction of economic activity in the East. The West German economy is growing faster than it has in nearly 15 years. Real GNP in the West has grown 4.6 percent in 1990 with 2.5 percent real growth forecast for 1991. In comparison, industrial production in the new federal states has plummeted—from August to October 1990, it had fallen 51 percent below the monthly levels of the previous year.

Ex ante, it should have been obvious that German economic unification would initially constitute a strong negative shock to the aggregate demand for East German goods and a positive shock to the aggregate demand for West German goods. East Germany went from a position of semi-autarky to one of complete opening to West Germany and all of West Germany's trading partners, and the average quality of East German products was far below the western standard. Without subsidies for many of the

products exported to Western markets (Pentacon cameras, for example), there was no reason to expect the initial impact of unification to increase East German "exports" to West Germany much, if at all. In fact, compared to one year earlier, September deliveries to East Germany had increased by 277 percent, while the flow of goods from East Germany only increased by 36 percent (*Handelsblatt*, November 26, 1990).

Despite the collapse of production in the new federal states, the economic and social union of East and West Germany has increased the disposable income of East German households. Unfortunately for East German producers, Western products have been getting a large share of that spending. Purchases by East German consumers contributed between one-half to a full percent of the real GNP growth in the *old* federal states in 1990, and are expected to account for another 0.5 percent of the growth in 1991.

The expected combined budget deficit in 1991 for all German federal, state, and local governments (including social insurance funds), on a national accounting basis, has been forecast by the Kiel Institute of World Economics at 130 billion DM. Estimates of the 1991 deficit of 150 billion DM (5.4 percent of GNP) are now reported in the financial press. Relative deficits of this size were last experienced in 1975 and 1982 when the government budget deficits were 5.7 and 4.1 percent of GNP, respectively.

Ideally, progress on deficit reduction over the next several years will come primarily from spending cuts, nontax sources of revenue (for example, private financing of infrastructure in the East and the privatization of government assets in the West) and economic growth in the new federal states.

The five leading West German economics research institutes estimate potential cuts in the defense budget of 5–6 billion DM and 10 billion DM in personnel and nonpersonnel expenditures, respectively (from a total defense budget of 55 billion DM). There are approximately 130 billion DM of government subsidies, including hidden tax breaks, that could be pruned back for sav-

ings of 20–30 billion DM annually. Privatization of a part of the telecommunications sector controlled by the Bundespost or of other federally owned firms such as Lufthansa is being considered as another possible source for deficit reduction.

Over the past decade, West Berlin has cost West German taxpayers about 20 billion DM annually. Between 1951 and 1990, the support for Berlin and the zone-border regions plus miscellaneous "costs of division," such as the payments for East German political prisoners, totalled over 400 billion DM, 85 percent of which went to West Berlin (*Handelsblatt*, November 12, 1990).

IV. The Political Economy of German Economic Integration

We close by reminding the reader of an obvious difference between 1948 and 1990—1948 is part of the historical experience shared in the united Germany of 1990. Erhard's success inspired the self-confidence of West German policymakers last spring as they pushed for a swift and radical break from the institutions of the centrally planned economy in East Germany.

In the East, we observe an understandable though unrealistic expectation that West German levels of real consumption can be achieved without closing the productivity gap between East and West. East German impatience for West German prosperity has been enhanced by an entitlement mentality that has survived the economic system that fostered it. These psychological factors still have the potential to complicate both the privatization and real adjustment processes of German economic integration. If in response to political demands, government intervention comes to dominate market adjustments during economic integration, structural change in the new federal states could likely repeat West Germany's experience with sectoral policy for declining industries on a much larger scale. The risk of the new federal states turning into a Mezzogiorno is genuine.

Will the new united German polity allow the market a chance to prove itself once

again? The economic question for Germany in the 1990's will be whether the German federal government can step back from playing the lead role in German integration now that the institutional foundations have been completed (pending successful privatization), and play only a supporting role in rebuilding East Germany's infrastructure and environment. In 1991, the Treuhand will no longer guarantee credit for working capital and most of the short-time workers will make the statistical transition to the ranks of the unemployed. Headlines will be dominated by news of plant closings and soaring unemployment. The demand for government action will be difficult to resist. German policymakers will need steady nerves and stamina if the benefits from the windfall of German economic integration are not to be squandered.

REFERENCES

- Breit, William, and Roger W. Spencer, *Lives of the Laureates, Seven Nobel Economists*, Cambridge: MIT Press, 1986.
- Collier, Irwin L., "The GDR Five-Year Plan 1986-1990," *Comparative Economic Studies*, Summer 1987, 29, 39-53.
- , "The Measurement and Interpretation of Real Consumption and Purchasing Power Parity for a Quantity Constrained Economy: The Case of East and West Germany," *Economica*, February 1989, 56, 109-20.
- Siebert, Horst, "The Economic Integration of Germany—An Update," *Kiel Discussion Papers*, No. 160a, September 1990.
- , "The Economic Integration of Germany: Real Economic Adjustment," *European Economic Review*, forthcoming.

CHINESE ECONOMIC REFORMS, 1979-89: LESSONS FOR THE FUTURE[†]

Chinese Enterprise Behavior Under the Reforms

By ROGER H. GORDON AND WEI LI*

During the recent decade of economic reforms from 1979 to 1989, the Chinese government adopted a series of policy and institutional changes aimed at increasing the productivity of the economy. While the initial efforts focused on agriculture, the government also tried to improve the performance of the nonagricultural sectors. Prior to the reforms, these sectors were heavily dominated by state-owned firms operating under central planning. The reforms decentralized many decisions to the firm level, or at least to the local government level. By 1985 most firms, for example, could change output quantity and variety, production technology, and the timing of production. To improve incentives, firms were allowed to retain a much larger fraction of profits. While the central planning apparatus remained in place, planned inputs and outputs became increasingly small fractions of the total inputs and outputs of firms. In principle they should not have affected marginal incentives. Outside-of-plan inputs and outputs could be traded freely among firms. While prices for goods allocated through the plan were tightly controlled,

there were increasing attempts to relax control over the prices at which trade took place outside of the plan. This type of dual-pricing system directly imitated the structure of the "responsibility system," which had proved so successful in agriculture.

Little attempt was made, however, to introduce factor markets in the state-enterprise sector. Even at the end of the period, firms continued to complain about being compelled to employ excess workers and workers could not quit without permission, though firms were given the discretion to hire temporary workers under contract. While the People's Construction Bank was set up to handle credit for new capital investment, proposed projects still needed to be approved at some level of government, and in practice investment decisions continued to be made primarily by the government planning ministries or by local governments.

Another key aspect of the economic reforms was to make it easier for local governments, collectives, and even private households to set up their own firms outside of the state planning structure. There was rapid growth in the number and output of such firms in response to the new flexibility. For example, the output of township firms grew dramatically from 8.6 percent of national nonagricultural output in 1979 to 26.6 percent in 1988. (See State Statistics Bureau, 1989, pp. 44; 248.)

The objective of this paper is to examine more closely the incentives created by these reforms, in order to better understand the causes of their success and failure. We start by discussing the state-owned sector, then examine the private and collective sectors.

[†]*Discussants:* Lawrence J. Lau, Stanford University; Richard D. Portes, Birbeck College; Gershon Feder, World Bank.

*Department of Economics, University of Michigan, Ann Arbor, MI 48109. We thank Michael Orszag, Michelle White, and Zhong Zhang for comments on an earlier draft. Some of the results reported in this paper come from a survey of 403 Chinese state enterprises conducted as part of a joint research project with the Chinese Academy of Social Sciences, funded by the Ford Foundation.

I. State-Owned Enterprises

Prior to the economic reforms, managers of state-owned enterprises in China were rewarded primarily based on their success in meeting physical output targets set by the government. With the reforms, rewards were instead based much more heavily on the firms' tax payments and accounting profits. While the tax rate on firms remained high by Western standards, in practice more than three-quarters of retained profits went to the firms' managers and workers. The link between accounting profits and manager/employee compensation might if anything have been stronger than in Western firms. These strong financial incentives, along with the decentralization of decision making, in theory should have led to a sizable increase in efficiency.

The story was not quite so simple, however. Given the large differences between accounting prices and market prices, accounting profits could well provide a poor approximation to true economic profits. When existing goods prices differ from market prices and allocation decisions are decentralized, barter trade should commonly arise at implicit prices for each type of good, which together are sufficient to clear all markets. What impact does barter trade have on firms' accounting profits and marginal incentives? If underpriced output is exchanged for a combination of cash plus a sufficient amount of underpriced input, then the two errors in the accounts should just offset, leaving pretax accounting profits unaffected. However, firms often sell underpriced output for underpriced consumer goods, or even for under-the-table cash payments. According to the evidence reported in our 1990 paper, sellers appear not to report receiving these extra payments, while buyers do report making them. The most compelling evidence was that firms reported that the inflation rate in the output price of machinery and industrial materials was very low, consistent with sales at official prices, whereas firms reported that the prices they paid when buying machinery and industrial materials grew at a rate roughly approximat-

ing the inflation rate in market prices. Since the responsibility for auditing a firm's accounts often rests with the firm itself, this evasion should not be surprising. This tax evasion increases the fraction of economic profits retained by the firm, reinforcing the incentives to raise efficiency. The amount of such evasion increased quickly during the period as market prices grew with inflation while official prices remained relatively constant.

The resulting loss in tax revenue led the government in 1986-87 to shift to a contract system in which each firm signed a contract with the government, typically for 5 years, agreeing to minimum tax payments during each year of the contract. But firms could still evade any tax on sales and profits above that necessary to justify the minimum tax payments, and in any case the inflation rate was high enough that real tax payments continued to fall.¹

Barter trade was not possible in all markets, however. Unless barter is costless, distorted official prices lead directly to distorted incentives. Often these distortions were sufficiently large that the government continued to intervene, either directly ordering decisions or else offering enough cheap credit or subsidized inputs to make the decisions it desired profitable from the firm's perspective. For example, the pay structure within firms remained highly egalitarian, so not surprisingly firms reported having an excess number of unskilled workers and a shortage of skilled workers. Given this pressure, the government continued to assign workers to firms and virtually forbade layoffs.

The interest rate on bank loans was also kept far below a market-clearing level. Not only was the interest rate low, but both interest *and* principal payments were nor-

¹In a sample of 361 firms with available data (see our 1990 paper), the annual growth rates of total tax payments from 1984 to 1987 were 10.48, 3.31, 1.47, and -1.35 percent, respectively, while the official annual retail price inflation in the same years were 2.8, 8.8, 6.0, and 7.3 percent, respectively. See State Statistics Bureau, p. 689.

mally tax deductible. Allowing these deductions is equivalent to allowing expensing of debt-financed new investments for tax purposes. In theory, under expensing no tax is collected in present value on a marginal investment—the cash flow to the government is proportional to that received by the firm, so when the firm breaks even so does the government. Therefore as a result of the below-market interest rate, debt-financed investments were subsidized on net.² Demand for loans exceeded the supply of funds, and increasingly so during the period as the rising inflation rate caused the real interest rate on loans to drop. Government intervention was necessary to ration credit.

When making allocation decisions, the government could not rely on the information available in a firm's financial accounts to judge the merit of any given project. That left the government with the option either of correcting the accounts for the distortions created by mispricing, or of continuing to plan directly the desired allocation of resources. Given the importance of the distortions, most importantly in interest rates, energy prices, foreign exchange rates, and the cost of skilled workers, and the pervasive way that these distortions filtered through the entire pricing structure, correcting the financial accounts would be virtually impossible.³ It is no wonder that the government was loathe to rely on such accounting figures to determine which firms to push into bankruptcy.

In spite of the problems with the financial accounting figures, they were still used heavily in determining the allocation of new investment, if only by requiring firms to invest a given fraction of their net-of-tax accounting profits. These reinvestment requirements were not self-enforcing, how-

ever—given the high tax rates on future profits, a firm's employees should much prefer the money now to receiving a fraction of the flow of net-of-tax profits resulting from investing the money in new capital. In practice, reinvestment rates were far below the levels required by regulation.

Even basing pay on a correct measure of economic profits would still create anomalous incentives if employees do not remain with the firm throughout the period affected by any given economic decision, or if employees fear a change in the compensation rule in the future. For example, managers might have an incentive to sell off existing capital in order to generate current profits and current bonuses, hoping that by the time the future losses show up they have been promoted or retired. Western firms try to avoid these incentives by basing the compensation of managers on changes in share values rather than on current profits. Chinese officials did recognize this problem. One experimental solution, known as the Asset Management Responsibility System, involved auctioning the job of manager every 5 years. The departing manager would receive a bonus based on the bid price of the new manager, thereby tying his compensation to the future prospects of the firm.

What happened in response to all these policy changes? In our earlier paper, we found in a sample of 403 state-owned enterprises that total factor productivity in the sample grew between 1983 and 1987 at a rate of 4 percent per year, while real output grew on average at 8 percent per year. Thus, decentralized decision making, even with badly distorted incentives, raised productivity in the state sector.

II. The Private and Collective Sector

The planning system in China, similar to that in the Soviet Union, favored heavy industry and highly capital-intensive technologies, resulting in a very different allocation of resources than is seen in market economies. For example, in 1986, 5 percent of workers were employed in the commercial trade sector in China, compared with 15 percent in Indonesia, 11 percent in Brazil,

²In fact, depreciation deductions were allowed as well, but during much of the period 40–50 percent of these deductions had to be handed over to the government, thereby mostly offsetting the gain from the tax deduction.

³See, for example, the complicated discussion in I. M. D. Little and James Mirrlees (1974) concerning how to correct appropriately for such distortions in cost-benefit analyses.

and 21 percent in the United States. (See State Statistics Bureau, p. 997.) Therefore, it should not be surprising that opening the economy to the entry of new firms would lead to rapid changes.

Private and collective enterprises responded swiftly to the economic liberalization. The former quickly expanded in the retail sector, while the latter became much more important in the industrial sector.⁴ Not only did these firms produce badly needed goods and services, but they also introduced competitive pressures into the economy, undermining the monopoly power of state-owned firms.

While the proliferation of private firms should not be surprising to Western economists, one should keep in mind that this occurred despite arbitrary taxes and erratic government interference, limiting, for example, their scope of trade, employment, and access to credit and scarce resources. Until 1985, for example, private firms could employ no more than seven workers. They also faced the risk of being shut down by the government at any point. Nonetheless, they grew and prospered.

The rapid growth of collectives, especially township enterprises, cannot be explained so simply. The local government normally controlled the access of collectives to workers, scarce inputs, and credits, and had to approve the appointment of managers, investment decisions, and even production plans. In addition, however, the local government kept most of the taxes paid by collectives in its jurisdiction,⁵ and had some control over the allocation of retained earnings. Since the local government effectively

owned and controlled these firms, and operated in competition with many other local jurisdictions, the resulting allocation decisions should have been relatively efficient. Each local government was small enough to take as given the implicit market prices for factors as well as goods—mobility across jurisdictions increased substantially during this period. The local government, by increasing efficiency, increased the resources under its control. Local government incentives remained distorted, however, to the degree to which some of the taxes paid by collectives had to be transferred to higher levels of government. In contrast, the national government's allocation decisions for state-owned firms would not be so constrained by market pressures. Thus, decentralization, even within the government, brought increased prosperity to the economy.

These improved incentives required decentralizing both control of tax revenue and control of the operations of firms to the same level of government. This was done not only for collectives but also for many smaller state-owned firms, implying the same improvement in incentives. Occasionally, however, control of the operations of larger state-owned firms was also delegated to local governments even though the bulk of the tax revenue went to the national government. The high national tax rates on these firms discouraged local investment in them, particularly when they were competing against locally owned firms whose tax revenue stayed within the jurisdiction.

Due to local taxes and distorted prices, locally controlled firms themselves faced different incentives than the local governments. As a result, local governments felt compelled to intervene regularly. This intervention often took the form of restrictions on trade with other jurisdictions. For example, when the implicit market price for a good exceeded the local opportunity cost of producing it, then the local government had an incentive to force local firms to buy the locally produced good, even if its price exceeded the implicit market price. Similarly, when the consumer price of a good exceeded its implicit market price, the local

⁴The market share of retail sales by private enterprises rose from 0.24 percent in 1979 to 17.8 percent in 1988, while the share of industrial output of the collectives rose from 21.53 percent in 1979 to 36.38 percent in 1988. See State Statistics Bureau, pp. 601; 267.

⁵Under the fiscal system adopted in 1985, all profit taxes from locally controlled firms including all collectives and some state firms go to the local government treasury, while other tax payments, such as the product tax and the value-added tax, were still shared with the national government.

government had an incentive to prevent outside firms and individuals from selling directly to local consumers, to keep the profits from the price differential within the community.

III. Lessons for the Future

The shift towards decentralized decision making during the reform period in China resulted in a rapid increase in output and in productivity. Yet in spite of these successes, the economic situation was very bleak by the end of the reform period. The problems arose mainly from the growing inflation rate. Official prices did not keep up with inflation, resulting in increasing price distortions and growing opportunities for tax evasion. The right to buy at official prices became increasingly valuable, resulting in growing corruption and growing rent-seeking behav-

ior, which undermined political support for the reforms. In addition, the increasing distortions to the measure of financial profits made oversight by the government of firm behavior, and even of the behavior of the economy as a whole, more and more tenuous.

REFERENCES

- Gordon, Roger H. and Wei Li, "The Change in Productivity of Chinese State Enterprises, 1983-1987," mimeo., University of Michigan, 1990.
- Little, I. M. D. and James Mirrlees, *Project Appraisal and Planning for Developing Countries*. London: Heinemann, 1974.
- State Statistics Bureau, *China Statistics Yearbook 1989*. Beijing: China Statistics Press, 1989.

Why Has Economic Reform Led to Inflation?

By BARRY NAUGHTON*

In China, as in other former command economies, inflation has threatened the economic reform process. Does inflation inevitably accompany the transition from command economic systems? Why did China (long regarded as the most successful example of gradualist reform) stumble into a severe inflationary episode at the end of the 1980's? This paper examines the Chinese experience of inflation, stressing the importance of fiscal imbalances.¹

The transition from a command to a market economy requires extensive change in the price system and in relative prices. In the command economy, prices were relatively unimportant in individual resource allocation decisions. Prices shaped some household consumption decisions, but most production and investment decisions were made by the government without reference to relative prices or profitability. Prices were fixed for long periods of time, and price relationships were highly distorted. The transition to a market economy requires that prices approach scarcity values, and play the dominant role in allocating resources in production and investment as well as consumption. Extensive changes in relative prices will inevitably be a part of an economic reform process, but must this lead to inflation?

The Chinese record, examined in Section I, shows that during the first 8 years of the reform process (1979–86), increases in the level of consumer prices were caused primarily by changes in relative prices. Food prices, long set at artificially low levels, were allowed to rise, and this accounted for most of the increase in the overall price level.

Beginning in 1987, however, prices of all important consumer goods began to increase, and inflation accelerated. Pure price level increases (the monetary component of price changes) became quantitatively more important than changes in relative prices. It was after this steadily worsening outbreak of inflation that tough deflationary policies were adopted that halted inflation, but also slowed growth and undermined reform (see my 1990 article).

Why did a period of moderate increases in the overall price level combined with substantial realignment of consumer prices give way to a period of pure inflation? To answer this question we must first reconsider the change in the function of the price system during the transition from a command economy. In the command economy, prices did not play a primary role in resource allocation, but the price system did serve as the primary instrument of resource mobilization. Government controlled prices were structured to create high prices and profitability for modern industry, at the cost of correspondingly low prices in agriculture, mining, and most services. This highly distorted price structure ensured that the bulk of national saving occurred in the state-owned modern industrial sector, the financial surpluses of which were then drawn into the government budget. Besides funding normal government services, the budget accounted for the bulk of national investment. Thus, investment was not guided by price signals; instead, the resources for government investment were provided by the overall price structure. Product price controls substituted for a taxation system, facilitating a large and intrusive government sector.

Clearly, if the price system is to serve its new function as the primary determinant of resource allocation, it must also shed its previous function as the primary instrument of resource mobilization. A single instrument cannot adequately perform two dif-

*University of California-San Diego, IR/PS, La Jolla, CA 92093.

¹This paper makes extensive use of official Chinese data, reprocessed into categories consistent with Western economic analysis. An appendix providing details of data sources and manipulation is available from the author on request.

ferent functions. As prices approach scarcity values during an economic reform process (and in particular as intersectoral price differentials diminish), macroeconomic stability requires either the creation of a new taxation system to fund government revenues, or the shrinkage of government outlays as revenues decline. If neither of these occur, or if they occur too slowly, government deficits will create inflationary pressures, the magnitude of which will depend on the way deficits are financed. Section II below documents the huge changes that have occurred in the role of the Chinese government budget during reforms. Overwhelmingly dependence on modern industry for revenues has characterized the budget throughout, but relative price changes in the economy have sharply reduced the profitability of modern industry. The result has been a huge decline in the share of budget revenues in GNP, a chronic fiscal crisis, and large overall public sector borrowing requirements. This has created a source of inflationary pressure separate from that created by the direct effect of relative price changes.

Fiscal deficits and price realignments cause inflation only if the banking system provides sufficient liquidity to accommodate price increases. Section III shows that monetary policy shifted to an extremely accommodative stance at the end of 1984. This shift was the final element that caused the gradual buildup and explosion of inflation in 1988. Inflationary pressures were temporarily contained by running a large deficit on international payments, but subsequently the deficit was narrowed while credit policy remained slack, and open inflation increased. This shift in monetary policy can be associated with the need to finance government investment through the banking system.

Section IV summarizes the argument. China initiated a process of price realignment without carrying out complementary reforms of the tax and fiscal system. As a result, budgetary revenues declined rapidly, eventually exceeding the point where government officials were willing to cut control over real resources correspondingly. They then turned to the banking system to gain

access to financial resources, accommodating the budget deficit and causing inflation. This suggests that inflation in China was not inevitable, but rather was the result of flaws in the sequencing of economic reforms.

I

Since the initiation of reforms, consumer prices have increased sharply on three occasions: 1980, 1985, and 1988. On each of these occasions, inflation was led by increasing food prices. Table 1 shows annual increases in the consumer price index, non-staple foods, and the slowest-growing component of consumer prices. While the price of staple grains has remained heavily controlled and subsidized, consumer prices of other food products have increased rapidly. State food prices have been raised, and consumers purchase an increasing portion of food on the free market. In each of these three episodes, food price increases were initiated by raising state prices and compensating urban workers with wage increases. In each case, after the initial inflationary surge, inflation moderated, and some food prices actually declined briefly. Thus, while changes in relative prices caused inflationary pressures, the government was initially able to contain those pressures.

TABLE 1—PERCENTAGE INCREASE IN
CONSUMER PRICES

	(1)	(2)	(3)
1978	0.7	2.2	0.0
1979	1.9	3.5	-0.6
1980	7.5	14.1	-4.5
1981	2.5	3.2	-2.0
1982	2.0	-0.2	-2.9
1983	2.0	4.6	-2.2
1984	2.7	6.0	-0.1
1985	11.9	23.0	1.2
1986	7.0	8.3	0.7
1987	8.8	14.9	3.2
1988	20.7	31.1	12.9
1989	16.3	13.8	14.3
Jan-June 90	1.5	0.2	-

Note: Col. (1) = Urban CPI; Col. (2) = Nonstaple foods; Col. (3) = The component of urban consumption outlays with the lowest inflation rate.

In addition to the episodes of food price adjustment, there are two distinct periods in the changing price level. Overall inflation rates were much higher between 1985 and 1989 than previously. During the 1978-84 period, food price increases were partially offset by reductions in the price of manufactured consumer goods, shown in column 3 of Table 1. Beginning in 1985, the overall price level began to increase at a substantially faster pace. Nevertheless, for 2 years, food price increases continued to be the major component of consumer price inflation, and this was the period of maximum price realignment. Beginning in 1987, increases in the overall price level became more important. Finally, during the third quarter of 1988, accelerating inflation and an announcement of imminent further price reforms led to panic buying and hoarding. This pushed inflation rates temporarily up to annual rates above 50 percent during July and August.

In the 1978-86 period, increased prices of nonstaple and miscellaneous foods, making up 40 percent of urban household outlays, accounted for 78 percent of the total increase in the price level; during 1986-89, they accounted for only 28 percent of the total increase in the price level. Explaining Chinese inflation requires that this shift between two substantially different regimes be explained.

II

Before economic reform, modern industry carried out the vast majority of national financial saving. In 1978, modern industry accounted for 12 percent of the labor force, but it generated a net financial surplus (total profits plus taxes collected from enterprises) equal to 25.4 percent of GNP, accounting for nearly four-fifths of national financial saving. Moreover, only small sums were retained by the enterprises, so virtually all of this, 24.7 percent of GNP, was remitted to the fiscal authorities. Household financial saving was small, amounting to only 1.3 percent of GNP, and households paid virtually no direct taxes. Thus the modern industrial sector, predominantly state owned, was

the dominant source both of national saving and of fiscal revenues.

The extraordinarily prominent role of modern industry in economywide financial balances was the result of the peculiar price structure discussed above. Low state-set agricultural procurement prices were most important, but low interest and depreciation charges were also significant. In order to maintain profitability differentials, substantial barriers to entry were erected around industry. Economic reform substantially altered price relationships, through three channels: 1) Agricultural procurement prices were raised relative to industrial prices, redistributing income to rural residents. Although increased prices of staple grains were absorbed by government subsidies, price increases of other foods and agricultural raw materials were passed through to urban residents and processing enterprises. 2) Depreciation charges and interest rates were increased. 3) Barriers to entry in the manufacturing sector were lowered, and industrial production by rural cooperatives and individuals, as well as local governments operating off-budget, exploded. Rapid entry into the most profitable manufacturing sectors exerted downward pressure on prices and substantially lowered profitability in industry overall.

As a result, financial surpluses in modern industry have declined drastically and steadily as a proportion of GNP since 1978. By 1988, total profit and tax of modern industry had declined to 16.7 percent of GNP, of which 3 percent of GNP was now retained by the enterprises, and 2 percent used to repay bank loans. Thus, industrial remittances to the budget declined from 24.7 to 11.6 percent of GNP, or by 13.1 percentage points. Budgetary revenues also declined drastically and steadily, by an amount almost exactly equal to the decline in industrial remittances. Between 1978 and 1988, budgetary revenues, appropriately adjusted and including all subsidies, declined from 35.4 to 19.8 percent of GNP, or 15.6 percentage points. Eighty-four percent of the decline in budgetary revenues is attributable to declining remittances from modern industry. There are few examples of

a nation's budgetary share declining by such a huge amount in such a short time (Mario Blejer and George Szapary, 1990).

It is worth emphasizing that the decline in the financial surplus of modern industry is not due either to slow growth of industry or to declining efficiency in the industrial sector. Industrial output has grown substantially more rapidly than GNP, and the proportion of the labor force in modern industry has increased to 15 percent of the total. The best efforts to measure total factor productivity in state industry, after correcting for price changes, shows significant increases through the period under consideration (Kuan Chen et al., 1988). Thus, the decline in profitability seems to be entirely attributable to the impact of changing relative prices, initiated by increased agricultural procurement prices and intensified by the downward pressure on relative manufacturing prices created by lowered barriers to entry.

To cope with the huge reduction in budgetary revenues, the Chinese government slashed expenditures, and the overt budgetary deficit has remained modest. Budgetary outlays declined from 35.4 percent of GNP in 1978 to 22.1 percent in 1988, yielding a budgetary deficit of 2.3 percent of GNP in 1988. The reduction in budgetary expenditure was accomplished by dramatic reductions in military expenditures and investment outlays. Government subsidies of all kinds (predominantly food grain) grew initially, to 7–8 percent of GNP, but as prices of nonstaple foods were raised, subsidies were stabilized at 6 percent of GNP. Yet these changes in the budget do not fully display the impact of the government on the economy. Budgetary financing of fixed investment declined sharply, but the central government investment plan did not decline in corresponding fashion. Completed investment under the central plan equaled 9 percent of GNP in 1978, but after initial reductions, recovered to 8 percent of GNP during the late 1980's. Taking into account the need to finance all stages of the investment process, less than half of the central investment plan was funded by budgetary revenues. The central government needed to borrow about 5 percent of GNP in order to

finance its investment plan. If we take the central government investment plan as a measure of the real use of investment resources mandated by the central government, the total government borrowing requirements is equal to the sum of the budgetary deficit and the deficit in the central investment plan, amounting to a total of 7 percent of GNP in 1988. A total deficit of this magnitude puts significant pressure on macroeconomic balance: beginning in 1987, total government borrowing requirements exceeded 50 percent of total credit creation by the banking system.

III

Changes in the structure of national saving have caused changes in the nature of the banking system and credit policy. Household saving has exploded, and annual household accumulation of monetary assets has surpassed 8 percent of GNP annually since 1984. As household saving has increased, the banking system has increasingly played the intermediary role common in most countries. Unlike the prereform pattern, the bank now channels saving from the household sector to businesses (still primarily state owned). Increased household saving made it possible to finance large public sector deficits under certain conditions.

Overall, credit policy has shifted sharply during different periods. Total lending grew between 1978 and 1983 at a 13 percent annual rate. From 1984, credit creation accelerated, and the rate of increase was sustained at 26 percent annually through 1987. This dramatic shift in credit policy immediately put upward pressure on prices, but these were dampened in 1985 by a trade deficit of 5 percent of GNP. Moreover, the output response was vigorous and industrial production grew rapidly. In subsequent years, though, as the trade deficit was gradually eliminated (by mid-1988), and inflationary expectations became rooted, the continuing expansionary credit policy led to an acceleration of inflation.

The abrupt shift to an expansionary credit policy contrasts with the gradual increase in total central government borrowing requirements. As central government requirements

increased in the mid-1980's, they threatened to crowd out the new decentralized investments that were fueling China's economic growth. At this point, the government, unwilling either to reduce its own borrowing or to crowd out decentralized investment, instead shifted to a much more expansionary credit policy. This shift was the proximate cause of inflation.

IV

The stylized facts presented above can be fully explained by two fundamental characteristics of the Chinese reform process. First, substantial progress was made in liberalizing product prices and introducing market forces into the economy. In particular, lower entry barriers in industry generated explosive growth in small-scale and rural output. Second, conversely, virtually no progress was made in rationalizing financial relationships, restructuring the tax system, and clarifying ownership rights and obligations for use of assets. The lack of a modernized tax system meant that changes in relative prices introduced in the wake of reforms caused a substantial erosion in government fiscal revenues. This might not be bad if the government were to acquiesce in a reduction of its control of resources. However, as the Chinese leadership sought to maintain a large central government investment program, they ordered the banking system to provide financing, and this led to the inflationary credit policy that ultimately disrupted the economy and the economic reform process.

This analysis suggests there is no inevitable reason why a gradual reform process must lead to severe inflation. Adjustment of relative prices does cause "spikes" of short-term inflation, but these can be managed if fiscal and monetary policy are sound. Instead, the cause of inflation is found in the sequencing of Chinese reform measures. If the Chinese government had carried out tax reform at an earlier stage of reforms, and broadened the tax base beyond the industrial sector, they might have been able to avoid the inflation of the 1980's. Gradualist reform strategies thus require early enactment of fiscal restructuring, but

there is nothing in the macroeconomic record that suggests the gradualist approach is inherently flawed.

Differences in economic system and level of development may limit the applicability of these lessons to European socialist countries. Having carried out economic reforms since 1978, China has a larger flex-price sector than any European command economy, with market prices existing alongside plan prices for nearly all commodities. Moreover, China's reforms achieved early success in agriculture, so that rapid growth of food output reduced the difficulty of price adjustment. Evidence of suppressed inflationary pressures is much greater in European command economies, especially the Soviet Union, than in China, where inflationary pressures can be openly expressed in free markets parallel to state controlled markets (but see Richard Portes and Anita Santorum, 1987). On the other side, the European socialist countries have greater administrative capabilities and, since 1989, fewer ideological constraints than China. As a result, they will certainly adopt reform strategies less gradual than that of China in the 1980's. However, all the socialist countries face difficult choices about the sequence and pacing of reforms, and the Chinese experience provides valuable information about the benefits and pitfalls of a gradualist approach.

REFERENCES

- Blejer, Mario and Szapary, George, "The Evolving Role of Tax Policy in China," *Journal of Comparative Economics*, September 1990, 14, 452-72.
- Chen, Kuan et al., "Productivity Change in Chinese Industry: 1953-85," *Journal of Comparative Economics*, December 1988, 12, 570-91.
- Naughton, Barry, "Economic Reform and the Chinese Political Crisis of 1989," *Journal of Asian Economics*, No. 2, 1990, 1, 349-61.
- Portes, Richard and Santorum, Anita, "Money and the Consumption Goods Market in China," *Journal of Comparative Economics*, September 1987, 11, 354-71.

Economic Reform of the Distribution Sector in China

By RICHARD H. HOLTON AND TERRY SICULAR*

Economic reform in China, slowed after the Tianamen Square incident in 1989, may be resuming. The Chinese Academy of Social Sciences has been asked by the central government to suggest further economic reforms in the distribution sector. Here we discuss this topic, noting first the role of distribution in the economic development process, then the changes in China's goods distribution process since economic reform was instituted. We conclude with observations about an ideal distribution system for China.

I. Distribution in the Process of Economic Development

The role of goods distribution in furthering economic development has received scant attention.¹ Yet distribution is clearly part of the production process when the latter is viewed in its entirety. As goods move from the producer to the final buyer through the channels of distribution, a variety of services must be performed ("produced") if final buyer demand is to be satisfied. Marketing textbooks note that the channels of distribution create form, time, place, and possession utilities; the channels provide information flows in both directions; assembling, grading, packaging, financing, risk taking, and negotiation are other components of the marketing process (Philip Kotler, 1980).

Improvements in technology, management, or organization can reduce the cost of production of distribution services, just as innovation in production can reduce the

cost of producing the physical good itself. Technological and managerial changes that reduce distribution costs affect the supply function for the product faced by final buyers. Assuming perfect competition at all levels, cost-reducing innovations in distribution will shift the supply function in the final buyers' market to the right, reducing the equilibrium price paid by end-users.

In an open economy, improvements in distribution might permit exports where none were possible before, even if the cost of producing the physical product were to remain unchanged. Thus a latent comparative advantage might be converted to an operative one.

It follows that increased efficiency in goods distribution can accelerate economic development just as can increased efficiency in the production of the goods themselves. Distribution might even be considered a candidate for a leading sector in the economic development process, that is, incremental investments in distribution might yield a higher net social return than investments in other sectors. Investment in distribution may also be a key during the economic transformation of socialist economies. Its importance is evident in the USSR, where inefficiency in the distribution system has caused food shortages even in years of bumper harvests.

II. Goods Distribution in China: An Overview

Before 1979, goods distribution in China was marked by much inefficiency. The physical infrastructure was primitive, units were overstaffed, inventories of unwanted goods accumulated, high-quality goods were scarce, and waste was palpable. Most consumer goods moved either through the Supply and Marketing Cooperatives or through the wholesale and retail system under the jurisdiction of the Ministry of Commerce. Grain was sold by the collectives or individ-

* University of California-Berkeley, Berkeley, CA 94720, and Harvard University, Cambridge, MA 02138, respectively. We thank the Ford Foundation for supporting this research.

¹ Unless otherwise noted, the term "distribution" herein will refer to the distribution of goods in the economy, not to income distribution.

ual growers to procurement stands under the Bureau (or Ministry) of Grain.² Capital equipment and other industrial products typically moved directly from producer to final buyer according to the dictates of the planning system. Planned distribution permeated the system, and the units in the system were assigned both customers and suppliers. Competition was absent.

A significant characteristic of the planning system was regional protectionism. Municipal authorities required the local department store to buy from the local wholesale company, for example, which in turn bought from local factories. Supplies from elsewhere were brought in only as a last resort, even though comparable quality goods might be available at lower prices (Bei Tao and Holton, 1989).

Since 1979, economic reforms have shattered the vertical, closed commercial system and allowed diverse channels to develop. Three aspects of these reforms are noteworthy. First, private individuals, collectives, producers, and government units outside the Ministry of Commerce can now buy and sell more freely. Second, the vast majority of small-scale, state-owned commercial enterprises have been converted to private or collective ownership, and efforts are being made to convert the rural supply and marketing cooperatives (that had in fact been state operated) to true cooperatives. Larger units in the state commercial system, while still state-owned, have been established as companies separate from the administrative bureaucracy. They are now managed under responsibility-system contracts that allow them to retain some profits, and they have more freedom to buy and sell as they wish. Third, the scope of planning has been greatly reduced, so that by the mid-1980's only a small number of commodities were subject to planned allocation at planned prices. By 1988 only half of China's retail sales took place at state-set prices (Almanac of China's Prices Editorial Committee, 1989, p. 351).

²At times this agency was an independent ministry; it is currently under the Ministry of Commerce.

A. Distribution of TV Sets

China's production of TV sets has exploded; the 1988 output, almost 25 million sets, was more than 50 times that of 1978 (State Statistical Bureau, 1990, p. 455). In 1985 there were 154 TV set producers in China, compared with 60 in 1978 (State Statistical Bureau, 1988, p. 251). Many of the new factories were set up by provincial and local governments outside the traditional production centers. Thus Shanghai, Beijing, and Tianjin, which in 1978 had produced 68 percent of China's TV sets, produced only 31 percent in 1987. (See State Council Office of the Leadership Group...1986, p. 249; State Statistical Bureau DITMS, 1988, p. 230.)

Prior to the reforms, the wholesale distribution of TV sets was monopolized by what are now the Metals, Transport, Electrical and Chemical Companies (MTECCs) under the Ministry of Commerce.³ The MTECCs purchased TV sets from factories and sold them to designated retailers. TV sets were in short supply and could only be purchased with ration coupons. State planning governed the quantities distributed and prices paid.

The MTECCs no longer dominate the distribution of TV sets. Manufacturers can now sell directly to consumers, as can private businesses, collectives and government departments outside the Ministry of Commerce. Furthermore the MTECCs at different levels and in different localities are no longer constrained and can compete with one another.⁴

The government continues to plan the production, distribution, and prices of TV sets, but the rapid growth in production has reduced the role of planning. Rationing has

³These companies handle distribution of bicycles and motorcycles, electrical appliances, small metal products such as hand tools, and household chemical products such as paint and cosmetics.

⁴Indeed, the employees of the Chengdu Municipal MTECC, formerly under the Sichuan Province MTECC, boasted that they outsell their former superior organization. (Interviews, Chengdu Municipal MTECC, July, 1990.)

been abandoned, and prices of black and white TV sets have been set free. For color TV sets, the price bureaus set retail prices and retail and wholesale margins. State-owned retail stores typically honor the planned retail prices, but private retailers frequently do not. Both state and private wholesalers often ignore the planned margins. During the panic buying of 1987–88 when market prices for TV sets rose rapidly, factories could earn higher profits by selling directly to consumers. They therefore refused to sell at the low planned factory prices. The state commercial sector, which continued to sell at planned retail prices, suffered.⁵

Early in 1989, the central government attempted to regain control of TV set distribution and to slow price increases by permitting TV sets to be sold only by agents with special permits. This shrank private wholesaling activity, but competition among the government departments responsible for issuing permits led to the overissue of permits, and so price increases continued.⁶

The demand for TV sets fell in 1989 and producers resumed selling at the planned factory price, but higher-quality models, still in short supply, continued to be sold directly by manufacturers at market prices. To sustain factories in the face of weaker demand, the MTECCs were required to buy certain quantities of TV sets at the planned prices. Yet the MTECCs and state retail stores were prohibited from stimulating sales by reducing prices, so inventories accumulated; the carrying costs reduced the MTECCs' profits. The MTECCs and state retailers have expanded advertising, extended warranties, and offered free gifts to buyers. Some have violated the regulations and reduced retail prices.

B. Grain Distribution

Since the 1950's, the Chinese government has tried to assure the urban sector of ade-

quate grain at low and stable prices. The rural population was expected to produce grain sufficient to supply its own needs and also to deliver planned quantities to the state at low prices. The margins between these planned purchase prices and the selling prices were small or negative, and so the government lost money on grain distribution. Interprovincial trade was small, averaging only 1.1 percent of domestic production in the late 1970's (Nicholas Lardy, 1990). Farm households and collectives could store and process grain for their own consumption, but the Grain Bureau had a virtual monopoly on grain marketing, storage and processing; private distribution of grain was minimal.⁷

Reforms since 1977 have freed grain trade considerably. Farmers can trade grain on free markets once they have fulfilled their delivery quotas; individuals can transport grain across county and provincial borders; state-owned units outside the Grain Bureau can trade and process grain. Retail grain markets have emerged, as have wholesale markets at traditional transshipment points such as Wuxi, Wuhu, and Shashi (Lardy). Planned grain procurement has been reduced and by 1989 only 70 percent of the grain marketed by farmers was sold to the Grain Bureau. And about half of the Grain Bureau purchases were at market-based "negotiated" prices. Thus by 1989 more than half of the grain marketed moved at market or market-based prices (Almanac of China's Commerce Editorial Committee, 1988, 1990; State Statistical Bureau, 1990).

This picture, however, probably exaggerates the development of freer markets for grain. "Negotiated" prices in some circumstances are set administratively and do not follow market trends. In urban areas, grain marketing is dominated by state units distributing at planned, ration prices, and the free market, serving only a limited demand, is sensitive to the planned price. In major grain-producing areas for rice (in the South), wheat (North), and corn (North and North-

⁵Interviews with MTECCs, retailers and producers in Sichuan and Beijing.

⁶Interviews, Chengdu and Chongqing, July, 1990.

⁷For a detailed discussion of prereform grain distribution, see Sicilar (1988).

west), the government dominates and strictly controls local marketing of the major grain. To assure that quotas will be filled, free markets are closed during procurement season and interregional trade is prohibited. When market trade is permitted, local governments often enforce price ceilings or influence the market price through their own market operations. In the case of rice, the central government has reimposed the Grain Bureau's monopoly; individuals and units outside the Grain Bureau cannot supply rice.

Competition among government units, common in the distribution of TV sets, is very limited in the case of grain and is discouraged by the authorities. For example, in the late 1980's, Guangdong reduced its planned grain distribution and allowed farmers to specialize in commercial crops. To procure grain, Guangdong's grain agencies bought in neighboring provinces. This raised prices in these provinces, and caused their grain agencies difficulties in enforcing quotas. Their complaints led the central government to prohibit Guangdong from buying grain in neighboring provinces.

Losses on grain distribution discourage competition among government units. The planned retail price is lower than the planned procurement price, and the central government reimburses the grain departments for only a portion of the resulting losses. Hence agents under the Grain Bureau have little incentive to seek new sources of grain or customers for grain. This situation motivates local governments to prohibit outsiders from coming in to buy grain; the outside demand would drive up the local price, put upward pressure on the government's prices, and make it more difficult to enforce quotas.

Competition from individuals and units outside the Grain Bureau is also limited. Producers can now market their grain themselves, but they are unable to mount a serious challenge to the state distribution system. Producers are numerous and small, with limited access to credit and long-distance transportation. The monopoly power of the Grain Bureau, however, is limited because farmers can resist selling to the

Grain Bureau when the terms are too unfavorable.

III. Conclusions

What features would characterize an "ideal" distribution system for China, and how can China move closer to such a system? One with perfect competition at all levels, from producers through the channels of distribution to final buyers, is clearly impossible. All the competitors at the wholesale or the retail level cannot have identical locations, so their offerings will be differentiated in at least this one dimension. Furthermore one can ask whether perfect competition is in fact desirable. In retailing, for example, consumers presumably want some variety in the products and services offered and so should be willing to pay slightly higher prices in order to enjoy that variety. Thus some degree of monopolistic competition may be consistent with the maximization of economic welfare (F. M. Scherer, 1980, p. 24).

This suggests that the ideal distribution system would be characterized by monopolistic competition, with large numbers of agents, some product and service differentiation, free entry, and no excess profits. But in many areas in China, markets are thin and thus oligopoly conditions are likely to apply to much wholesale and retail trade if the units are to be of an efficient size. One option would be simply to let the free market work and to tolerate the resulting quasi-monopoly profits. These profits might be held in check if competition or the threat of entry were enhanced by improving transportation and providing buyers with good information about prices in adjacent markets.

An alternative approach would be to recognize that many enterprises in distribution are natural monopolies and hence should be regulated as are public utilities in the United States. Restricted entry would permit each unit to operate at a point approximating the least cost rate of output. Price controls would apply. The prereform system in China approximated this model, except that the units were government operated.

As noted earlier, the government-owned and regulated approach to the design of the distribution system has not been considered satisfactory in China.

For some products the current dual pricing system could be a compromise solution, that is, part of the output is sold at the state plan price and the remainder at market prices or at prices approximating the free market. But if planned and market prices differ, illegal arbitrage and rent-seeking behavior emerge. Producers find excuses for not delivering the planned output at the state price, since sales at market prices are more profitable. Moreover, as the experience with TV sets and grain demonstrates, the dual price system tends to break down. For any particular product, either the plan is effectively eroded by the market, or the government's efforts to enforce the plan limit the marketing channel and prevent it from operating efficiently.

A second and perhaps superior compromise policy would let the free market operate, but for certain products the government might impose price ceilings or floors. In effect this would limit the economic rents earned by quasi monopolists in distribution. Thus the government could limit the monopsony power of grain merchants in rural localities by setting a minimum price to be paid to farmers. At the retail level monopoly rents on basic necessities could be limited by mandatory price ceilings.

One could argue that this policy at the retail level would not be effective because the quasi monopolist, prohibited from charging a profit-maximizing price for a basic food item, for example, would simply charge more for other products in the product mix. But presumably the profit-maximizing merchant is already pricing each product at a point where marginal cost equals marginal revenue; the profit-maximizing price on item A would not be affected by the price charged for item B. A second objection is that price ceilings and floors can be difficult and costly to maintain. Successful cases do exist, however; witness the Indonesian price policy for rice (C. Peter Timmer, 1989).

Developments in the distribution of TV sets and grain in China suggest the condi-

tions likely to promote movement toward the ideal distribution system. Success is more likely when 1) some players exist who are powerful enough to challenge the state distribution system, 2) agencies within the state distribution system have incentives to compete among themselves, and 3) production is growing rapidly.

In the case of TV sets, producers were in a position to challenge the local MTECCs because any one locality generally contained a single, relatively large TV set manufacturer. This producer had access to bank credit, trucks, and other resources needed for marketing. Since TV set producers generated revenues for the local government (local governments taxed the profits of the local manufacturing units) and were usually closely linked with a local government bureau, they also had political bargaining power. Such conditions did not hold for grain producers.

TV set distribution has also been characterized by competition among branches and departments under the Ministry of Commerce. This competition has arisen because, following managerial reforms establishing them as profit-earning companies separate from the Ministry's administrative bureaucracy, these agencies have realized that they can earn profits from competitive behavior. Departments under the Grain Bureau, however, earn small or negative profits from handling grain and hence are not motivated to handle more.

Finally, rapid growth in the production of TV sets has caused TV set distribution to outgrow government planning. Once a surplus of TV sets emerged in the late 1980's, government efforts to contain competition eased. In the case of grain, output grew rapidly in the early 1980's and surpluses emerged temporarily in 1983-84. At that time the government proposed reforms substantially liberalizing grain distribution, but these proposals were largely abandoned when output fell in 1985 and stagnated thereafter (see Sicular, 1990). Persisting fear of grain shortages has since hindered further reforms in grain distribution.

These considerations imply that simply allowing private trade and competition to emerge will not necessarily lead the econ-

omy toward the ideal distribution system. Progress toward the ideal system is likely to differ among markets and, since government commercial departments will try to maintain their dominance, depends critically on price and management reforms that encourage such agencies to compete among themselves. Meanwhile, improvements in information and transportation will lower barriers to entry and promote a more efficient distribution system.

REFERENCES

- Kotler, Philip, *Principles of Marketing*, Englewood Cliffs: Prentice-Hall, 1980.
- Lardy, Nicholas, *China's Interprovincial Grain Marketing and Import Demand*, U.S.D.A. Economic Research Service, Agriculture and Trade Analysis Division, Staff Report No. AGES 9059, Washington: USGPO, 1990.
- Scherer, F. M., *Industrial Market Structure and Economic Performance*, 2nd ed., Chicago: Rand McNally, 1980.
- Sicular, Terry, "Grain Pricing: A Key Link in Chinese Economic Policy," *Modern China*, October 1988, 14, 451-86.
- _____, "Ten Years of Reform: Progress and Setbacks in China's Agricultural Planning and Pricing," Harvard Institute for Economic Research, Discussion Paper No. 1474, March 1990.
- Tao, Bei and Holton, Richard H., "Interprovincial Trade and Economic Development in China," *China Economic Review*, Spring 1989, 1, 23-32.
- Timmer, C. Peter, "Indonesia: Transition from Food Importer to Exporter," in T. Sicular, ed., *Food Price Policy in Asia: A Comparative Study*, Ithaca: Cornell University Press, 1989.
- Almanac of China's Commerce Editorial Committee, *Zhongguo Shangye Nianjian, 1988* (1990), Beijing: Zhongguo Shangye Chubanshe, 1988; 1990.
- Almanac of China's Prices Editorial Committee, *Zhongguo Wujia Nianjian, 1989*, Beijing: Zhongguo Wujia Chubanshe, 1989.
- State Council Office of the Leadership Group for the National Industrial Census and State Statistical Bureau Department of Industrial, Transportation and Materials Statistics, *Zhongguo Gongye Jingji Tongji Ziliao 1986*, Beijing: Zhongguo Tongji Chubanshe, 1986.
- State Statistical Bureau, Department of Industrial, Transportation and Materials Statistics (DITMS), *Zhongguo Gongye Jingji Tongji Nianjian, 1988*, Beijing: Zhongguo Tongji Chubanshe, 1988.
- _____, *Statistical Yearbook of China, 1988*, Hong Kong: Economic Information & Agency, 1988.
- _____, *Zhongguo Tongji Nianjian, 1990*, Beijing: Zhongguo Tongji Chubanshe, 1990.

Shareholder Heterogeneity: Evidence and Implications

By LAURIE SIMON BAGWELL*

The perfect market paradigm provides a powerful foundation for financial theory. In perfect capital markets, there are no transaction costs, all traders have equal and costless access to information, and traders act as price takers. If existing claims "span" the state space, excess supply curves are perfectly elastic. Moreover, differences in preferences or beliefs do not result in disagreement among shareholders about firm policies. Underlying this unanimity is the shared valuation of the stock, which translates into agreement about firm strategies. The ability to transact without affecting the market price is central to many important propositions, including the Modigliani-Miller irrelevance theorems.

This paper examines the nature of supply curves for corporate equity. Until recently there has been little direct empirical assessment of their elasticity. At issue is whether or not the supposition of shareholder homogeneity of valuations (and its implications) represents a good approximation to actual markets. This paper's call for further empirical evaluation of shareholder valuations echoes the perspective offered by Eugene Fama and Merton Miller, who in discussing perfect markets observed that

[N]o such market exists in the real world, nor could it. Rather, what we have here is an idealization...permit[ing] us to focus more sharply on a

limited number of aspects of the problem and usually greatly facilitat[ing] both the derivation and statement of the sought-for empirical generalizations. In the nature of the case, however, the generalizations so obtained can never be anything more than approximations to the real phenomena that they are supposed to represent. The question is whether, considered as approximations, they are close enough; and this, of course, is a question that can only be answered empirically and in light of the specific uses to which the approximations are put.

[1972, pp. 21-22]

This paper provides evidence that current shareholders' valuations differ *dramatically*. This provocative empirical finding implies that the hypothesis of common valuations indeed is not always a good approximation. If the approximation is poor, then conclusions stemming from it must be reconsidered. This requires additional analysis of the microeconomic foundations of disagreement in shareholder valuations, and the contexts where shareholder disagreement is substantial.

I. Evidence: Supply Curve Elasticity

My earlier paper (1990a), investigating the extent to which the supply curves for equity deviate from perfect elasticity, examines shareholder tendering responses in Dutch auction repurchases of stock. In Dutch auctions, the company states the number of shares it will repurchase, and a price range within which stockholders can offer to sell their shares. Shareholders fill out tendering schedules indicating how many shares they are willing to sell at each price within this range. It is a dominant strategy for atomistic

[†]*Discussants:* Susan Collins, Council of Economic Advisers; Howard Kunreuther, University of Pennsylvania; Colin Camerer, University of Pennsylvania.

*Department of Finance, Kellogg Graduate School of Management, Northwestern University, Evanston, IL 60208. Comments from Kyle Bagwell, David Brown, and Ken Judd, and support from NSF grant no. SES-8821666 are gratefully acknowledged.

shareholders to tender their shares at their true valuations.

The firm then compiles the tendering responses from the lowest to the highest price, constructing the supply curve for the stock. All stockholders who tendered at prices at or below the minimum price necessary to acquire the number of shares the company seeks receive the purchase price for their tendered shares. In the sample examined, on average, 16.7 percent of the outstanding shares are tendered at or below the purchase price, which is at a 13.4 percent premium above the preannouncement market price.

Firms are not required to disclose the shareholder tendering responses. However, 32 of the 52 firms conducting Dutch auctions between 1981 and 1988 disclosed this proprietary information to me. The individual tendering responses provide a unique opportunity to examine directly the elasticity of the supply curve for stock.

The supply curves documented in Dutch auction repurchases have a distinct upward slope. When bids are ranked from lowest to highest, the average difference between the 1st and 6th percentile bid is 4.4 percent of the preannouncement market price, from the 6th to 11th is 2.6 percent, and from the 11th to 16th is 2.0 percent. That is, the difference between the 16th percentile shareholder valuation and the 1st percentile shareholder valuation is 9.1 percent. The average arc elasticity of the supply curve is 1.67. Formal regression analysis confirms the significant upward slope of these curves.

Evidence consistent with upward-sloping supply curves is detected in similar transactions. M. Bradley et al. (1988) find that the premium paid in interfirm tender offers is increasing in the fraction of target shares purchased by the acquirer. David Brown and M. Ryngaert (1990) find results similar to Bradley et al.'s for fixed price repurchase tender offers.

Andrei Shleifer (1986) also provides evidence suggestive of an upward-sloping supply curve, finding that the share prices of firms added to the S&P 500 Index increase at the announcement of the inclusion. The magnitude of the price increase is positively

related to the increased buying of the shares by Index funds. Since being included does not signal any information about stock value, the findings suggest that the price increase is being driven by increased demand in the presence of an upward-sloping supply curve.

II. Supply Curve Elasticity and Information Interpretations

In light of the evidence suggestive of less than perfect elasticity for stock, I reexamine the traditional interpretations of the share price reaction to specific corporate events. In perfect capital markets, the number of shares traded in a given stock has no effect on its price. If the market is less than perfect, a large purchase (sale) of shares could inflate (depress) the price of the shares temporarily due to market illiquidity. Further, the number of shares traded can carry new information about the stock that would cause a permanent reassessment of share value. While these alternative hypotheses have been considered extensively in the existing literature, few papers have allowed for a third possibility: a large purchase (sale) of shares could inflate (depress) the price of the shares permanently due to an upward-sloping supply curve. If the excess supply curve is less than perfectly elastic, then a large purchase (sale) alters the marginal holder of stock to one with a higher (lower) reservation price.

The typical Dutch auction repurchase buys 15 percent of the outstanding shares and increases the market price at its announcement by 7.8 percent. These price increases are frequently attributed to new information. This interpretation is appropriate when the supply curve is perfectly elastic; if the supply curve is flat, only new information can change the stock price. However, attributing the price change solely to information is misleading when the supply curve is upward sloping because, in addition to any shift in the supply curve, movement along the supply curve is confounded with new information. Specifically, if we assume that the repurchase conveyed no information, an elasticity estimate of 1.67 would nevertheless imply an average price

increase of 9.1 percent, exceeding the observed announcement effect.

Myron Scholes (1972) finds a permanent negative price reaction to the sale of large blocks. Since there is a permanent price effect, he concludes that the sale signals information to other traders. Greater price changes occur if the seller is presumed to have adverse information motivating the sale. Wayne Mikkelson and M. Megan Partch (1985) reconsider block sales in light of an upward-sloping supply curve. They document a significant negative price reaction to block sales regardless of the type of seller. Further, the magnitude of the price response is positively related to the size of the offering. While this work suggests possibly important lasting supply effects, they find no relationship between the price reaction and elasticity determinants they consider. Thus, further empirical analysis is needed to determine the relative importance of the information, liquidity and supply components of the price reaction, for not only block trades but all changes in the supply or demand of shares.

III. Supply Curve Elasticity and Corporate Control

Under shareholder unanimity, the composition of a firm's shareholders, and the elaborate rules governing shareholder voting, have little impact. Under this supposition it is difficult to explain why these rules vary dramatically across firms, and why changes in these rules result in large changes in stock prices. An example of the insights gained by allowing for shareholder heterogeneity is the role it plays in the choice of cash distribution method.

Cash distributions are usually explained as ways to signal information, alter leverage, or disgorge free cash, but few theories have untangled the choice between alternative methods of distribution. One plausible explanation is that the method of distribution is influenced by its effects on the nature of the shareholder population. For example, my article (1991) argues that distributing cash through a share repurchase as opposed to dividends serves as an effective takeover deterrent in the presence of an upward-

sloping supply curve for stock. Shareholders willing to tender in a repurchase are systematically those with the lowest valuations. The repurchase therefore skews the distribution of the remaining shareholders towards a more expensive pool, raising the cost of a takeover to the acquirer. Interestingly, targets of takeover activity account for nearly half of all recent repurchases.

On the other hand, my paper with Kenneth Judd (1989) shows that dividends may be chosen even if they are tax disadvantaged relative to share repurchase, because dividends do not change the population of shareholders. This may be desirable since it maintains the current majority. Hence, while a repurchase deters a takeover by altering the shareholder population, dividends may be desirable precisely because they do not alter the population of shareholders in the absence of takeover concerns.

When shareholders have differing preferences and transactions are costly, sometimes opposed shareholders may choose to remain together within a firm, fighting over firm decisions, instead of incurring the costs of portfolio reshuffling. Since firm decisions are made in light of the conflict, shareholder nonunanimity makes corporate control decisions central. In particular, focusing on shareholder disagreement may help us evaluate the recent prevalence of supermajority requirements and other corporate charter amendments. Rene Stulz (1988) argues that supermajority rules effectively change the marginal shareholder to one requiring a higher premium, thereby benefiting shareholders while making takeovers less likely. Allowing for shareholder heterogeneity may also yield an understanding of the increased use of nontraditional takeover mechanisms like two-tiered tender offers, and the impact of legislation like that passed in Delaware in 1988, requiring that hostile takeovers be approved by 85 percent of all nonaligned shareholders.

IV. Conclusion

Despite the importance of common shareholder valuations to finance theory and practice, there is direct evidence of significant shareholder disagreement in Dutch

auction repurchases. It is vital that we learn more about the market conditions underlying shareholder heterogeneity and the contexts where they are significant.

Perhaps the most important unanswered question is: what causes the significant deviation from perfect elasticity? My paper (1990b) sheds light on this question by examining empirically the cross-sectional determinants of supply curve elasticity in Dutch auction repurchases. Preliminary findings indicate that supply curves are more elastic when institutional holdings are high, dividend yield is high, price has not varied much in the past 5 years, and the fraction bought back is large.

Supply curves may be more elastic when institutional holdings are high because institutions have small capital gains liabilities or institutional investors share consensus. Capital gains taxes induce shareholders with lower basis values to value the share more highly; many institutions are tax exempt. Further consistent with tax-induced heterogeneity, low tax bracket investors typically hold high dividend yield stocks, hence, we would observe more elastic supply curves for these stocks. Since price variability may result in increased dispersion of basis values, these supply curves would be less elastic. Many of the same variables affect the tendering rates for fixed price share repurchases in Brown and Ryngaert. These results demand further examination of nontax sources of heterogeneity, including asymmetric information and divergence of opinion.

Knowledge of the relative importance of taxes, transactions costs, or asymmetric information for shareholder disagreement affects tax reform and regulatory policies. Consider, for example, lowering the tax rate on capital gains. Since taxes on capital gains induce shareholders with different capital gains liabilities to value shares differently, decreasing the capital gains tax diminishes the disparity of liabilities across shareholders, increasing consensus. If taxation is an important source of shareholder heterogeneity, this change in policy could lead to increased agreement among investors. Similarly, if asymmetric information is an important source of shareholder heterogeneity,

then changes in regulatory policy for required disclosure of security trading may also have important implications for shareholder disagreement. These policies, though designed for taxation or regulation, might radically change the market for corporate control.

REFERENCES

- Bagwell, Laurie Simon, (1990a) "Dutch Auction Repurchases: An Analysis of Shareholder Heterogeneity," unpublished, June 1990.
- _____, (1990b) "The Sources of Shareholder Heterogeneity," unpublished, November 1990.
- _____, "Share Repurchase and Takeover Deterrence," *Rand Journal of Economics*, forthcoming, Spring 1991.
- _____, and Judd, Kenneth L., "Transaction Costs and Corporate Control," unpublished, December 1989.
- Bradley, M., Desai, A., and Kim, E. H., "Synergistic Gains from Corporate Acquisitions and their Division between the Stockholders of Target and Acquiring Firms," *Journal of Financial Economics*, No. 1, 1988, 21, 3-40.
- Brown, David and Ryngaert, M., "Heterogeneous Shareholders: Evidence from Buybacks and Control Contests," unpublished, 1990.
- Fama, Eugene F. and Miller, Merton H., *The Theory of Finance*, Hinsdale: Dryden Press, 1972.
- Mikkelson, Wayne H. and Partch, M. Megan, "Stock Price Effects and Costs of Secondary Distributions," *Journal of Financial Economics*, June 1985, 14, 165-94.
- Scholes, Myron, "The Market for Securities: Substitution versus Price Pressure and the Effects of Information on Share Prices," *Journal of Business*, April 1972, 45, 179-211.
- Shleifer, Andrei, "Do Demand Curves for Stock Slope Down?," *Journal of Finance*, No. 3, 1986, 41, 579-90.
- Stulz, Rene, "Managerial Control of Voting Rights: Financing Policies and the Market for Corporate Control," *Journal of Financial Economics*, No. 1/2, 1988, 20, 25-54.

Investor Diversification and International Equity Markets

By KENNETH R. FRENCH AND JAMES M. POTERBA*

Since the fortunes of different nations do not always move together, investors can diversify their portfolios by holding assets in several countries. The benefits of international diversification have been recognized for decades. In spite of this, most investors hold nearly all of their wealth in domestic assets. In this paper we use a simple model of investor preferences and behavior to show that current portfolio patterns imply that investors in each nation expect returns in their domestic equity market to be several hundred basis points higher than returns in other markets. The lack of diversification appears to be the result of investor choices, rather than institutional constraints.

I. International Asset Ownership Patterns

Most corporate equity is held by domestic investors. The domestic ownership shares of the world's five largest stock markets are: United States, 92.2 percent; Japan, 95.7 percent; United Kingdom, 92 percent; Germany, 79 percent; and France, 89.4 percent. This information, and other data on cross-border equity transactions, can be used to estimate the international equity holdings of investors in each country. Table 1 presents crude estimates of the equity portfolio allocation for investors in the United States, United Kingdom, and Japan.¹

*Graduate School of Business, University of Chicago, Chicago, IL 60637, and Department of Economics, MIT, Cambridge, MA 02139, respectively. We are grateful to the NSF, CRSP, the Alfred P. Sloan Foundation, and the John M. Olin Foundation for research support, to Michael Howell and Vincent Koen for data assistance, and to Cole Kendall, Richard Thaler, and Richard Zeckhauser for helpful comments. A data appendix for this project is available from the ICPSR in Ann Arbor, MI, or from the authors. This paper is part of the NBER program on Financial Markets and Monetary Economics.

¹These estimates cumulate the net purchases of equity by investors in each country, with adjustments

TABLE 1—EQUITY PORTFOLIO WEIGHTS:
BRITISH, JAPANESE, U.S. INVESTORS

	Portfolio Weight			Adj. Market Value
	U.S.	Japan	U.K.	
U.S.	938	.0131	.059	\$2941.3
Japan	031	.9811	.048	1632.9
U.K.	011	.0019	.820	849.8
France	005	.0013	.032	265.4
Germany	005	.0013	.035	235.8
Canada	010	.0012	.006	233.5

Note: Estimates correspond to portfolio holdings in December, 1989. They are based on the authors' tabulations using data from the U.S. *Treasury Bulletin* and Michael Howell and Angela Cozzini (1990). Adjusted market values exclude intercorporate cross-holdings from total market value, and correspond to June 1990 values.

The estimates show little cross-border diversification for U.S. and Japanese investors. At the end of 1989, Japanese investors had only 1.9 percent of their equity in foreign stocks, while U.S. investors held 6.2 percent of their equity portfolio overseas. The British, by comparison, held 18 percent of their portfolio abroad, divided almost equally among the United States, continental Europe, and Japan.

Since the United Kingdom is a smaller share of the total world equity market than the United States or Japan, it is not surprising that its investors hold more equity outside their own borders. However, the diversification of U.K. portfolios is a recent phenomenon. At the end of 1979, U.K. pension funds, which today hold 21 percent of their assets in foreign equities, held only 6 percent of their portfolios abroad (Michael

for both stock market and exchange rate movements. Ian Cooper and Costas Kaplanis (1986) also estimate cross-border equity holdings, but their calculations are largely imputations that do not rely on country-by-country equity flows.

Howell and Angela Cozzini, 1990, p. 30). The growth of international equity investments followed Prime Minister Thatcher's relaxation of capital controls.

II. Is Incomplete Diversification Costly?

The gains from diversification depend on the correlation of returns in different equity markets. We compute real returns from the perspective of a U.S. investor, assuming the investor uses three-month forward contracts to lock in an exchange rate for the amount of his initial investment each quarter. The average pairwise correlation between quarterly returns on the equity markets in the United States, Japan, the United Kingdom, France, Germany, and Canada for the 1975–89 period is .502. This suggests that nontrivial risk reduction is available from cross-border holdings. The correlations are similar if the returns are measured in yen or pounds, and whether or not the exchange rate risk is hedged.

To measure the costs of incomplete diversification, we assume that a representative investor in each country has a constant relative-risk-aversion utility function defined over wealth, $U(W) = -e^{-\lambda W/W_0}$, and that he maximizes expected utility.² For a given set of portfolio weights w associated with a vector of mean returns μ and a covariance matrix Σ , this implies an expected utility of

$$(1) \quad E[U(w)] = -e^{-\lambda(w\mu - \lambda w'\Sigma w/2)}.$$

In this setting, optimal portfolio weights w^* satisfy

$$(2) \quad \mu = \lambda w^{*'} \Sigma.$$

With limited historical data on international equity returns, it is difficult to measure expected returns, μ , or to infer the optimal portfolio weights, w^* , with any precision. We can, however, make reasonable estimates of the covariance matrix, Σ . Under the assumption that investors put all

their wealth in the equity of the six largest stock markets, we can ask what set of expected returns, $\mu^*(w, \Sigma)$, would explain the pattern of international portfolio holdings we observe. We use equation (2) to calculate the expected returns implied by the actual portfolio holdings of U.S., Japanese, and British investors. We also compute the expected returns implied by an international "value-weighted" portfolio strategy for investors in each nation. The last column of Table 1 shows the value weights, based on market capitalization data from Morgan Stanley Capital International but with corrections for intercorporate equity holdings as in our earlier article. The adjustment reduces the importance of the Japanese and German markets.

Panel A of Table 2 shows that substantial differences in expected returns *across countries* for investors in a given nation are needed to rationalize observed portfolio holdings. In the most extreme case, British investors must expect annual returns in the U.K. market more than 500 basis points above those in the U.S. market to explain their 82 percent investment in domestic shares. This large implied differential reflects the substantially higher standard deviation of returns on the British market, relative to returns on the U.S. and Japanese markets. For U.S. investors, the annual expected return on U.S. stocks must be 250 basis points above the expected return on Japanese stocks. In contrast, for Japanese investors, the expected return on Japanese stocks must be 350 basis points above the expected return on U.S. stocks.

The difference in expectations for different investors judging the same market are also striking. Our estimates suggest that Japanese investors, for example, expect returns from Japanese stocks which are more than 300 basis points greater than the returns U.S. investors expect. There are similar differences in the expectations of foreign and domestic investors in both the U.S. and U.K. equity markets.

Although these differences in expected returns are striking, the implied alternative of equal expected returns across all markets may not be an appropriate benchmark. As

²We set $\lambda = 3$; see our 1990 paper for more detail on calibration.

TABLE 2—EXPECTED REAL RETURNS
IMPLIED BY ACTUAL PORTFOLIO HOLDINGS

	U.S.	Japan	U.K.
A. Expected Returns Needed to Justify Observed Portfolio Weights			
U.S.	5.5	3.1	4.4
Japan	3.2	6.6	3.8
U.K.	4.5	3.8	9.6
France	4.3	3.4	5.3
Germany	3.6	3.0	4.8
Canada	4.7	3.0	4.0
B. Deviation Between Implied Returns for Actual and Value-Weighted Portfolios			
U.S.	0.9	-1.5	-0.2
Japan	-1.1	2.5	-0.3
U.K.	-0.7	-1.4	4.4
France	-0.3	-1.2	0.7
Germany	-0.2	-0.8	1.0
Canada	0.5	-1.2	-0.2

Note: See text for further description of calculations.

another alternative, we estimate the expected returns that would induce investors in each country to hold an international value-weighted stock portfolio. The *difference* between the expected return vector implied by each country's actual investment pattern, and that implied by a value-weighted strategy, is shown in Panel B of Table 2. The results again suggest that investors expect domestic returns that are systematically higher than those implied by a diversified portfolio. The differences between the two sets of implied returns for U.S. and British investors, however, are rarely larger than 100 basis points. For example, U.S. investors' concentrated holdings of U.S. stocks can be explained by "optimistic" expectations of roughly 90 basis points. A similar "pessimism" of about 110 basis points is needed to justify U.S. investors' underweighting of the Japanese market. Explaining the behavior of both Japanese and British investors requires more "optimism" regarding their own markets: 250 basis points for the Japanese, and over 400 basis points in the United Kingdom.

III. Institutional and Behavioral Explanations for Underdiversification

What explains the apparent tendency for portfolio investors, particularly in the United Kingdom and Japan, to overweight their

own equity market? There are two broad explanations. First, institutional factors may reduce returns from investing abroad or they may explicitly limit investors' ability to hold foreign stocks. It is difficult, however, to identify such constraints. Institutional barriers are unlikely to explain the low level of crossborder equity investment *today*, even though capital controls substantially restricted equity flows in the 1970's. Tax burdens that are higher on foreign than domestic equity income should lead investors toward holding domestic equity. There is little difference, however, between foreign and domestic tax burdens for most investors. Although all of the nations we examine impose a dividend withholding tax on payments to foreign shareholders, typically these payments can be credited against taxes in the investors' home country.³

Transaction costs also appear unable to explain limited international diversification. The cost of trading may be lower in more liquid markets such as New York than elsewhere, but this should incline all investors toward the most liquid market, not toward their own domestic market. Since all shares must be held by someone, differences in transaction costs should be reflected in differences in expected returns. The large gross equity flows across borders also suggest that transaction costs cannot explain why investors specialize in their home markets. For the United States in 1989, gross foreign equity purchases were fifty times net purchases (see our earlier paper).

Explicit limits on cross-border investment could also affect portfolio holdings, although few of them appear to bind at present. In France, for example, a foreign investor may not hold more than 20 percent of any firm without authorization from the Ministry of Economy and Finance. In Japan, insurance companies cannot hold more than 30 percent of their assets in foreign securities. Many U.S. pension funds traditionally

³Tax-exempt investors may face a burden from such taxes, since they have no tax liability against which to claim the credit. Even for these investors, however, the tax would only reduce expected after-tax returns in foreign markets by about 50 basis points.

interpreted the "prudent man" rule as limiting their degree of international exposure.

The current level of international portfolio investment seems to be well below any institutional constraints. In the mid-1980's, for example, foreign investors were substantial net sellers of Japanese shares. Similarly, foreigners were net sellers of U.S. equities in 1988. Such reductions in international equity investments suggest that constraints on foreign holding are not binding, implying that incomplete diversification is the result of investor choices.

A second class of explanations for imperfect diversification focuses on investor behavior. One important possibility is that return expectations vary systematically across groups of investors. Robert Shiller et al. (1990) report direct evidence on this question. In early 1990, they surveyed portfolio managers in Japan and the United States. The U.S. investors expected an average return of -0.3 percent on the Dow Jones Industrial Average over the next twelve months, compared with an expected return of -9.1 percent on the Nikkei. In contrast, Japanese investors expected an average return of 12.6 percent on the Dow, and 10.8 percent on the Nikkei. While the Japanese investors were more optimistic than their U.S. counterparts with respect to both markets, they were relatively more optimistic about the Tokyo market.

The statistical uncertainties associated with estimating expected returns in equity markets makes it difficult for investors to learn that expected returns in domestic markets are not systematically higher than those abroad. The standard error of the estimated mean annual return on the U.S. stock market, based on 60 years of data, is 200 basis points. Thus, the 95 percent confidence interval for the mean return spans 800 basis points. Because it is difficult to estimate *ex ante* returns, investors may follow their own idiosyncratic investment rules with impunity.

Another important behavioral insight concerns the perception of risk in equity markets. Investors may not evaluate the risk of different investments based solely on the historical standard deviation of returns. They may impute extra "risk" to foreign

investments because they know less about foreign markets, institutions, and firms.⁴ Country-specific closed-end mutual funds, popular in the United States during the late 1980's, may overcome these fears (see Catherine Bosner-Neal et al. 1990, for a discussion).

Although the level of cross-border equity investment is low, it is growing and with time the international diversification puzzle may recede. Cross-border equity investment patterns may nevertheless provide important insights on how investors value risk and how they select portfolios. The evidence of incomplete diversification presented here is consistent with evidence from many other markets. Ronald Lease et al. (1974) show that in the late 1960's, many individuals held relatively few stocks. Both the mean and median in their sample of investors were close to eleven different securities. The rise of index mutual funds in the last two decades has improved the diversification of individual investors, but directly held equity still accounts for two and one-half times as much of household wealth as *all* mutual funds, of which index funds are only a small part.

Perhaps the most striking example of incomplete diversification is the tendency of most households to own residential real estate near where they work. The returns on their human and physical capital may consequently be highly correlated. This generates a much less diversified portfolio than holding, for example, a real estate investment trust with a national real estate portfolio.

⁴Amos Tversky and Chip Heath (1991) present evidence that households behave as though unfamiliar gambles are riskier than familiar gambles, even when they assign identical probability distributions to the two gambles.

REFERENCES

- Bosner-Neal, Catherine et al., "International Investment Restrictions and Closed-End Country Fund Prices," *Journal of Finance*, June 1990, 45, 523-47.
- Cooper, Ian and Kaplanis, Costas, "Costs to

- Crossborder Investment and International Equity Market Equilibrium," in J. Edwards et al., eds., *Recent Advances in Corporate Finance*, Cambridge: Cambridge University Press, 1986.
- French, Kenneth R., and Poterba, James M., "Japanese and U.S. Cross-border Common Stock Investments," *Journal of the Japanese and International Economics*, December 1990, 4, 476-93.
- _____ and _____, "Were Japanese Stock Prices Too High?," *Journal of Financial Economics*, forthcoming, 1991.
- Howell, Michael and Cozzini, Angela, *International Equity Flows-1990 Edition*, London: Salomon Brothers European Equity Research, 1990.
- Lease, Ronald, Lewellen, Wilbur and Schlarbaum, Gary, "Individual Investor Attributes and Attitudes," *Journal of Finance*, May 1974, 29, 413-33.
- Shiller, Robert J., Fumiko Kon-ya and Yoshiro Tsutsui, "Speculative Behavior in the Stock Markets: Evidence from the U.S. & Japan," mimeo., Yale University, 1990.
- Tversky, Amos and Heath, Chip, "Preferences and Beliefs: Ambiguity and Competence in Choice Under Uncertainty," *Journal of Risk and Uncertainty*, January 1991, 4, 5-28.

Window Dressing By Pension Fund Managers

By

JOSEF LAKONISHOK, ANDREI SHLEIFER, RICHARD THALER, AND ROBERT VISHNY*

The conventional wisdom on Wall Street is that portfolio managers are reluctant to produce annual reports that show holdings of shares that have sharply declined in value. One money manager is quoted by S. Jansson as saying: "Nobody wants to be caught showing last quarter's disasters.... You throw out the duds because you don't want to have to apologize for and defend a stock's presence to clients even though your investment judgment may be to hold" (1983, p. 139). The practice of deleting such shares from portfolios at year end is an example of "window dressing." Despite conventional wisdom, little is known of how widespread this practice is. In this paper, we look for evidence of window dressing among an important group of portfolio managers, those who manage pension funds.

As of December 1989, pension funds owned close to \$900 billion of American equities, about 25 percent of the stock market capitalization. Of this amount, about \$700 billion is actively managed and the rest is invested in index funds. Some plans manage their money internally, but more commonly they hire several money managers and split the pension plan's money among them. Plan sponsors typically evaluate fund managers once a quarter, but the main evaluation takes place at the end of the year. Based on these evaluations, assets are reallocated across money managers. Window dressing, if it occurs, is presumably a response to these evaluations.

Why might a fund manager engage in window dressing? The main focus of fund

evaluations is past performance relative to some benchmark, such as the S&P 500. Stock returns, however, are noisy and may not suffice to identify the manager's investment philosophy or to see whether he just got unlucky on a few stocks. As a result, sponsors may look at the actual portfolio holdings as well. To impress sponsors, fund managers may alter these portfolios at the end of the quarter, and especially at the end of the year, that is, window dress. A finding that money managers window dress is therefore evidence that managers are evaluated on a broader set of criteria than performance alone.

Selling losers is the most frequently mentioned form of window dressing. Pension funds can also reduce the pace of sale of winners, increase the purchases of winners and reduce the purchases of losers to impress the sponsors with the looks of their portfolios.¹ Some of these strategies, however, may not be as effective as selling losers. When a fund buys winners after they rose in price, sponsors would realize that these winners were not held during the price rise. If a fund purchases losers, its manager can probably explain that he bought them after they had fallen. On the other hand, slowing down the sale of winners might be as attractive as speeding up the sale of losers. We will examine these less frequently mentioned and perhaps less sophisticated forms of window dressing as well as the generic dumping losers strategy.

*University of Illinois, Champaign, IL 61820; Harvard University, Cambridge, MA 02138; Cornell University, Ithaca NY 14850; and University of Chicago, Chicago, IL 60637; respectively. We are grateful to Gil Beebower and Vasant Kamath for help with this project.

¹R. Haugen and Lakonishok (1988) suggest window dressing as a possible explanation of the "January Effect." Lakonishok and S. Smidt (1988) find a very substantial increase in the Dow Jones Industrial Average (an index of 30 large stocks) in the last week of December. They offer window dressing as a possible explanation of this finding.

I. The Data

Our analysis is based on a proprietary sample of 769 equity pension funds provided by SEI, a professional fund evaluation service. Equity funds hold at least 90 percent of their money in equities. The data set contains, for each fund, complete equity portfolio holdings at the end of each quarter from 1985 to 1989. Sponsors of most of these funds are corporations, but there are some state, municipal, and endowment pension funds as well. The total amount under management in these 769 funds at the end of 1989 is \$124 billion, or about 18 percent of the total actively managed holdings of pension funds. The average equity holdings of a fund at the end of 1989 are \$161 million; the largest 5 percent of the funds manage 65 percent of the money.

We do not observe all the trades that a fund makes in a given quarter because a fund might have bought and sold the same stock within that quarter. However, we can estimate purchases and sales based on portfolio changes from quarter end to quarter end. In the sample, in an average quarter, a fund buys 13 percent of the dollar value of its previous quarter's holdings, and sells 12 percent. Purchases exceed sales in part because funds receive dividends and in part because new contributions to the funds from the sponsors typically exceed payouts to the retirees. The turnover rate of 50 percent a year is typical for institutions. Pension funds' holdings are heavily concentrated in the largest capitalization stocks. While the bottom 50 percent of stocks in terms of size comprise 4.3 percent of market capitalization, they only represent 1.4 percent of pension funds' equity holdings.

To determine whether managers engage in window dressing, we must classify stocks into performance categories. At the end of each quarter, we take the CRSP universe of all NYSE, American, and NASDAQ stocks, and divide it into equal size quintiles based on stock returns over *the past year up to the end of that quarter*. This procedure is repeated every quarter, so the composition of each quintile changes each quarter. Since we know which stocks belong to which quin-

TABLE 1—HOLDING, BUYING, SELLING BY PAST PERFORMANCE QUINTILE, ALL FUNDS COMBINED^a

	Past Performance Quintile				
	1	2	3	4	5
Universe Holdings in Quintile as % of Total Universe Holdings	.06	.16	.23	.28	.28
Fund Holdings in Quintile as % of Total Fund Holdings	.05	.17	.25	.29	.24
Sales in Quintile as % of Total Sales	.06	.17	.22	.27	.28
Purchases in Quintile as % of Total Purchases	.06	.19	.24	.28	.23

^aAll numbers are averaged over quarters and years.

tile, we can compute how much each fund holds, buys, and sells each quarter from each quintile. We use the average of beginning and end-of-quarter stock prices to estimate the total value of purchases and sales in each quintile by each fund.

Table 1 presents the fractions of their stock portfolios that funds as a whole hold in each past performance quintile (averaging over years and quarters) and the fraction of value of the CRSP universe in those quintiles. Table 1 also reports the fraction of purchases and sales that the funds as a whole make in each quintile. Holdings, sales, and purchases are all computed as if all money was in one aggregate fund. Table 1 shows that funds hold less of extreme losers and winners than exist in the universe. Funds hold only 5 percent of their portfolios in extreme losers, compared to 6 percent for the universe portfolio. They also hold 24 percent of their portfolios in extreme winners, compared to 28 percent for the universe. Obviously, funds are overrepresented in holdings of intermediate performance quintiles. One reason for the underrepresentation in extreme performance categories is that funds hold few small stocks, which are more likely to be among the extreme losers and extreme winners.

The evidence on purchases and sales reveals that trading is not always in proportion to holdings. Extreme losers represent 6 percent of both buying and selling, but just 5 percent of the holdings. Funds clearly *trade* extreme losers more. Perhaps they get rid of "mistakes," consistent with window dressing, but they also buy stocks they think are undervalued. Winners in contrast are overrepresented in sales relative to holdings and underrepresented in purchases relative to holdings. One way to summarize these data is that funds are generally contrarian (buy losers and sell winners) except they get rid of mistakes. The next section looks at the data more closely by comparing the fourth quarter to the first three.

II. Activity of Pension Funds by Quarter

All accounts of window dressing stress the relative importance of the year-end report. This section compares trading in the first three quarters with that in the fourth.

About 20 percent of the funds in an average quarter have zero holdings in the bottom performance quintile. To avoid statistical problems this raises when we examine the cross section of funds, we combine the bottom two quintiles, so the bottom group in a quarter is 40 percent worst performing stocks over the previous year. As a result of combining quintiles, funds holdings are more evenly distributed across the four groups.

Our measure of selling intensity in performance group i in quarter j by fund k is

$$(1) \quad \frac{\text{SELL}(i, j, k) / \text{HOLD}(i, j-1, k)}{\sum_i \text{SELL}(i, j, k) / \sum_i \text{HOLD}(i, j-1, k)}$$

where $\text{SELL}(i, j, k)$ is the value of sales by fund k in quarter j and group i , and $\text{HOLD}(i, j-1, k)$ is the value of holdings at the end of the previous quarter of the exact same stocks as those in performance group i in quarter j . Both SELL and HOLD are defined using the average of the beginning and end of quarter j prices. The numerator of (1) is the ratio of the sales in a performance group to holdings of the same stocks at the end of the previous quarter. The

denominator is the ratio of total sales in this quarter to holdings at the end of the previous quarter.

Equation (1) measures the selling intensity of a fund in a given performance group relative to the overall selling intensity. It corrects for the fact that funds might be selling less in a given group at a given time only because they are selling less of everything. If a fund sold 50 percent of the extreme winners it holds, but only 20 percent of its total holdings, its sales index for winners according to (1) is 2.5. Having computed this number for every fund, quarter and year, we average it over funds, years, and quarters 1-3, and then compare the average over quarters 1-3 to that for quarter 4.

Our measure of buying intensity by fund k in quarter j in performance group i is

$$(2) \quad \frac{\text{BUY}(i, j, k) / \sum_i \text{BUY}(i, j, k)}{\text{UNIV.HOLD}(i, j) / \sum_i \text{UNIV.HOLD}(i, j)}$$

where $\text{BUY}(i, j, k)$ is dollar purchases by fund k in group i in quarter j and $\text{UNIV.HOLD}(i, j)$ is the value of CRSP universe holdings in quarter j in group i . Again, both variables are computed using the average of beginning and end of quarter j prices. This number is the fraction of purchases that a fund makes in a performance category relative to the fraction of universe holdings in that performance category. So if 20 percent of a fund's purchases were in performance group 4, which is only 10 percent of the universe value in that quarter, the number for the fund in that quarter and quintile is 2.

Equation (2) controls for the value of each performance group that a fund can buy in every quarter, so purchases are scaled by the availability of stocks to buy. In this respect, the numbers for sales and purchases are parallel: the universe holdings define availability of stocks for purchase, and since funds in general do not take short positions, their own holdings define the availability of stocks for sale. We compute (2) for each fund, and then, as with measure

TABLE 2—SALES AND PURCHASES OF PENSION FUNDS BY QUARTER^a

Quintile	Quarter 1-3	Quarter 4	<i>t</i> -Test ^b
A. Sales Relative to Own Holdings			
1-2	1.32 (.026)	1.42 (.033)	1.75
3	.92 (.009)	1.02 (.084)	1.76
4	.91 (.007)	.89 (.014)	.24
5	1.44 (.053)	1.37 (.028)	1.22
B. Purchases Relative to Universe Holdings			
1-2	1.22 (.012)	1.33 (.024)	5.47
3	1.07 (.009)	1.09 (.017)	.097
4	1.01 (.008)	1.00 (.013)	.44
5	.90 (.009)	.88 (.015)	.71

^aSales number is from equation (1), purchase number is from equation (2). Both are averaged over funds and years.

^bFor difference, Quarter 1-3 = Quarter 4.

(1), average over funds and years for each quarter, and finally average over quarters 1 through 3 and compare that average to the one for quarter 4.

Table 2 presents the results, with cross-sectional standard errors for each quarter-performance group mean (in parentheses) and *t*-tests of equality of means for quarters 1-3 on the one hand and quarter 4 on the other. Unlike the results in Table 1 that were computed as if all holdings and trades were by one aggregate fund, in Table 2 we compute measures (1) and (2) first for each fund and then average over funds. This enables us to compute cross-sectional standard errors.

To begin, consider the findings for quarters 1-3. First, when it comes to purchases, funds are clearly contrarian: relative to availability they overbuy losers (ratio of 1.22) and underbuy winners (ratio of .90). Second, when it comes to sales funds oversell winners relative to their holdings (ratio of 1.44), but they also oversell losers (ratio of 1.32). Much less selling activity is observed in the middle quintile groups. These results confirm Table 1's findings that funds are

generally contrarian, except that they oversell extreme losers, consistent with the window-dressing hypothesis.

The comparison across quarters shows that the sale of losers relative to other quarters does accelerate in the fourth quarter, as it should if fund evaluations are most critical at the end of quarter four. In quarters 1-3, an average fund sells 32 percent more of the extreme losers (relative to its holdings of those losers) than it sells of all stocks; in quarter 4 this measure rises to 42 percent ($t = 1.75$). Consistent with window dressing, funds actively sell losers, especially in the fourth quarter.

What about other forms of window dressing? There is weak evidence of increased demand for winners in the fourth quarter: the sale of winners slows down ($t = 1.22$) although purchases do not rise. In addition, there is strong evidence of an increase in the purchases of losers in the fourth quarter relative to the first three ($t = 5.47$). In quarters 1-3, funds buy 22 percent more losers than they would if they bought stocks randomly from the universe, in quarter 4 this number rises to 33 percent. This result is robust in a variety of specifications. It is inconsistent with the simple version of window dressing, according to which just holding a portfolio that performed well (regardless of how long they have been held) is rewarded by the sponsors, but consistent with the more sophisticated view that buying losers after they have fallen can be justified to the sponsors. Funds might buy losers in the fourth quarter to provide liquidity for individuals who are selling them to realize losses in the current tax year. In this case, losers might be particularly good bargains in the fourth quarter.

To check robustness of these results, we have also performed the analysis by value-weighting the funds rather than equally weighting them. This amounts to assuming that all the funds' holdings and trades are done by one aggregate fund. The results, which are not presented, are qualitatively similar, although dumping losers is less pronounced when funds are value-weighted. For example, in quarters 1-3, value-weighted sales of losers relative to holdings

(corrected for all sales relative to holdings) are 1.0, but for quarter 4, this number is 1.06. This result suggests that window dressing is more prevalent in small funds. Indeed, we have performed the analysis for the 20 percent smallest funds as of the first quarter of 1985, and found a stronger propensity to sell losers in general, and in the fourth quarter in particular, for this subsample. For small funds, our measure of selling activity for losers is 1.37 in quarters 1–3 and 1.59 in quarter 4 ($t = 2.02$). More extensive window dressing by small funds suggests that it may be too hard to fool the sophisticated sponsors of large funds by window dressing, so their managers do not even try.

III. Conclusion

This paper took a preliminary look at portfolio strategies of a sample of 769 pension funds representing about 18 percent of private pension equity holdings in the United States. We have found that the average pension fund is contrarian, in that it buys more intensively stocks that have performed poorly and sells a disproportionately high fraction of stocks that have performed well. Such contrarian investment tendencies of the funds suggest that, contrary to a common perception, trading practices of institutions might reduce volatility of stock prices. This result obviously deserves a closer look.

We have also found evidence that in every quarter, funds sell poorly performing stocks disproportionately to their holdings,

that is, they get rid of “mistakes.” Moreover, the pace of dumping mistakes accelerates in the fourth quarter, consistent with the most common window-dressing strategy. This result is stronger for small funds. This finding supports the view that window dressing is a response to costly monitoring by fund sponsors of individual portfolio decisions, for it pays less for large funds, whose portfolios are monitored more frequently and completely, to pursue this practice.

Interestingly, purchases of losers also rise in the fourth quarter. This suggests that sponsors do not just look at the funds’ holdings of winners and losers, but recognize that buying losers after their prices fell might be smart. The overall evidence reveals some window dressing by pension fund managers. However, we stress that fourth-quarter strategies do not radically depart from the usual practice, but rather accentuate the typical investment strategy of funds: buy losers and sell both extreme winners and losers.

REFERENCES

- Haugen, R., and Lakonishok, J., *The Incredible January Effect*, Homewood: Dow Jones Irwin, 1988.
- Jansson, S., “The Fine Art of Window Dressing,” *Institutional Investor*, December 1983, 139–40.
- Lakonishok, J. and Smidt, S., “A Seasonal Anomalies Real: A Ninety Year Perspective,” *Review of Financial Studies*, Winter 1988, 1, 403–26.

The Rationality Struggle: Illustrations from Financial Markets

By JAYENDU PATEL, RICHARD ZECKHAUSER, AND DARRYLL HENDRICKS*

For most economists it is an article of faith that financial markets reach rational aggregate outcomes, despite the irrational behavior of some participants, since sophisticated players stand ready to capitalize on the mistakes of the naive. (This process, which we call poaching, includes but is not limited to arbitrage.) Yet financial markets have been subject to speculative fads, from Dutch tulip mania to junk bonds, and to occasional dramatic losses in value, such as occurred in October 1987, that are hard to interpret as rational. Descriptive decision theory, especially psychology (see D. Kahneman et al., 1982), can help to explain such aberrant macrophenomena. Here we propose some behavioral explanations of overall market outcomes—specifically of financial flows, that are of considerable practical consequence to both policymakers and finance practitioners.

I. Behavioral Explanations of Market Macrophenomena

Investors play for significant stakes and have sustained opportunities for practice—both factors that should promote rational outcomes. C. Camerer (1987) shows that even in experimental markets, practice and significant payment do away with many anomalies. Moreover, since nonrational investors lose their funds, natural selection operates. Yet overall outcomes still may deviate from rationality, depending on two factors: the rationality of individual participants and the opportunities for poaching. Table 1 outlines the possibilities.

When participants are mostly rational and there are many opportunities for arbitrage (cell 1), we expect the efficient markets

TABLE 1—MODELING OUTCOMES OF SOCIAL INTERACTIONS: POACHABILITY AND RATIONALITY

Rationality ^a	Poachability/Arbitrage Potential:	
	High	Low
Full/Substantial; Individualistic	1) Efficient Markets	2) Anomalies Due to Incomplete Use and Flow of Information
Bounded/Low; Possibly Relativistic	3) Natural Selection Processes—pressures to restore efficiency	4) Grossly Inefficient Outcomes

Notes: Cell 2: For example, prices of open-end mutual funds fail to reflect management practices and skills; Cell 3: For example, commodity and gambling markets; Cell 4: For example, misallocated individual portfolios, over- and undershooting by groups.

^aNature/Proportion; Orientation.

paradigm to triumph. In cells 2 and 3, results are best explained by a merger of behavioral considerations and economic analysis. In 2, we expect economic paradigms to succeed, albeit with behavioral residues associated with problems of information flows. Purchases of open-end mutual funds, that cannot be poached, might display some anomalies. In 3, barring new entrants, processes of natural selection reduce the role of nonrational players over time. In cell 4, behavioral models should provide important insights into inefficient outcomes. No one can benefit from (poach on) the misbalanced portfolio or poor retirement funding decisions of another. Mistakes are to be expected, though beyond a threshold they may provoke a corrective response. (A 1990 information campaign by Harvard led to a one-third increase in employees' use of highly tax-advantaged supplemental retirement annuities, which had long been available.)

Cell 4 offers other interesting possibilities when preferences focus on relative, not absolute, outcomes. Judging one's allocations relative to those of others (as in labor market contests) requires less information gathering. This approach may also be dictated by envy. Such relative valuations could lead

*John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138. Our research was supported by Decision, Risk, and Management Sciences, NSF.

decision makers to distort their own decisions, say in a "keep up with the Joneses" effort, or an attempt to move with the herd as a mechanism of protection.

We introduce two behavioral hypotheses to help explain financial phenomena: *Barn Door Closing* for mutual fund purchases, and *Herd Migration Behavior* for debt-equity ratios. Barn door closing, in the horse protection sense, refers to undertaking behavior today that would have been profitable yesterday. Herd migrations in finance occur when market conditions change, so that individual decision makers wish to alter their holdings substantially. Their transition is slowed because they seek protection by traveling with the herd.

II. Barn Door Closing—Purchases of Mutual Funds

The Nobel Prize-winning contributions of Markowitz and Sharpe address the rational portfolio choice problem and its implications for market pricing and efficient portfolio decisions. Their models imply that any individual's optimal mix of asset holdings will comprise a market portfolio (with assets in proportion to their total market value weights) and a riskless fund. When combined with the efficient markets hypothesis, this view leads to a passive (i.e., nontrading) portfolio strategy. The recently introduced multifactor Arbitrage Pricing Theory has similar implications, though the implied universal fund components remain to be satisfactorily identified. If all investors behaved as the financial models predict, observed flows should be due entirely to liquidity/consumption needs or incremental savings and should be explained by portfolio balance considerations.

In contrast, our earlier 1990 paper finds that relative flows across *individual* open-end equity mutual funds reflect 1) status quo bias, 2) a performance effect (i.e., investors' belief that a managed fund with a superior past will perform better than they could as individuals, a view encouraged by financial professionals), and 3) framing/data packaging. In cross-sectional time-series regressions (for 96 funds over 1975–87), we

explain 76 percent of the variance (R -squared) and highlight three interesting behavioral influences:

1) Persistence (Status quo bias). Other things equal, a one-dollar incremental flow induces a 75 cent flow in the following period. Although the avoidance of learning costs may justify some behavior persistence, we conjecture that investors (individuals more generally) shortchange important decisions by spreading their attention too evenly. Persistence may also spring from regret avoidance—an attempt to avoid a mistaken act of commission.

2) Past performance. A one-percentage-point return higher than the average fund's return implies a \$200,000 increased flow in the next year (where the median fund's size is \$80 million and the median flow is \$21 million).

3) Framing and data packaging. Rank measures, which are widely reported, appear to be more relevant in explaining flow patterns than are cardinal risk-adjusted performance measures.

In explaining net flows of funds from individuals to the mutual fund sector, barn door closing behavior may be relevant. Mutual fund purchasers may exhibit it because they: 1) rely on trends/patterns (widely prescribed for and practiced in commodities trading but contrary to efficient markets theory and near-martingale asset prices); or 2) engage in personal window dressing (realigning their portfolio to a desirable composition for the sample period experienced).

The traditional struggle of the multiple selves (see Thomas Schelling, 1982, pp. 57–82) is to control one's present self on behalf of the future self. We add a backward-looking feature. Individuals, seeking to contain regret, may try to remove reminders of their past errors. To invoke the agency framework, investors engaged in personal window dressing are imperfectly monitoring principals deceiving themselves into thinking better of their agents, namely their earlier selves.

If the barn door closing hypothesis is germane, individuals will buy more mutual funds after the stock market goes up, and sell after it plunges. Further, if there is

some behavioral threshold effect, this reaction will manifest itself mainly for large changes. Consider the fraction, f , of the U.S. household sector's flow of financial purchases (composed of direct and intermediated net purchases of equities, bonds, and short-maturity or demand deposits) directed to mutual funds. Quarterly data on f is constructed from the Federal Flow of Funds for the 1952Q1–1990Q1 period and excludes households' indirect claims, such as pensions. The fraction f appears mean stationary over the sample period, though during the 1980's f is higher (0.11) than the overall mean (0.05).

In a regression of f on four of its own lags, changes in Treasury bill interest rates, and returns on the equity market (proxied by the value-weighted NYSE index), we observe an economically large and statistically significant positive coefficient on equity returns that is consistent with barn door closing. (Simultaneity problems are avoided since most mutual funds are open end, and hence their supply of shares is highly elastic.) When the market has done well, one wishes that one had invested there, rather than purchase short-term assets (that normally represent more than 50 percent of the households' flows) and fixed-interest assets (that represent about 21 percent), and vice versa. A regression that decomposes equity returns into large ($> |10 \text{ percent}|$) and small changes suggests that barn door closing is most relevant when a threshold has been exceeded. The coefficient on large changes is 0.35 (t -statistic of 4.09), whereas the one on small changes is insignificant (0.06 with a t -statistic of 0.47). Thus a 1 percent change in equity returns beyond a threshold change of 10 percent induces a 6 percent change in the rate of investment in mutual funds by households. (The results are not driven by some subperiod; of the 25 large-change observations, 10 are from the 1950's and 1960's, and 15 are from the 1970's and 1980's.)

III. Financial Herd Migrations—Corporate Debt-Equity Ratios

Migrating birds and trekking wildebeest all know that traveling in a group offers protection. Financial players also may mi-

grate in herds, as when firms increase their debt-equity, S&Ls invest in junk bonds, and banks increase their Third-World debt holdings. These transitions are not instantaneous for many reasons, including the superior information aggregation and mutually informed choices that result from movements in clusters. As with our animal friends, it may be dangerous to get too far out of line.

Each decision maker in a financial migration balances the benefits of more quickly approaching the optimum against the costs of moving away from the herd. As each takes small steps, the whole process ratchets along. Financial migrations, unlike periodic animal migrations, tend to chart unfamiliar territories, and the optimal destination often is not clear. This uncertainty, combined with the natural tendency of individuals to free ride on the information of others, provides the potential for overshooting; as we saw with Third-World debt, or proceeding a while along the wrong path (like the wildebeest who plunge one after another into a ravine, none having had sufficient incentive to worry about his own direction). Birds, scientists now believe, are guided to their distant destinations by the stars and magnetic fields, through navigational methods buried deep in their genes. Human decision makers are less blessed, and must call on their brains.

Even the most clear-sighted financial navigator may be deterred from steering his own course if there is herding on the other side of the market. Banks have delayed writing down doomed real estate loans because the market "would not understand." Corporations considering an increase in their debt-equity ratio had to be concerned about the perceptions of lenders and investors, who might be unfamiliar with the Modigliani-Miller theorem. Bankers hesitate to lend to a firm whose debt-equity ratio tops its industry.

We examined the annual ratios of debt (book value) to equity (market value) for the 200 largest firms (by sales) during the period 1971–89. On the COMPUSTAT database, 15 of these firms had some missing observations, and 3 clearly had outlier debt-equity ratios (in excess of 5): this left us with a

usable sample of 182 firms. We assigned firms to 10 industries based on a reasonable classification of two-digit SIC codes. Over the sample period, there was a persistent overall rise in debt-equity ratios, though considerable heterogeneity across industries and firms (R. Taggart, 1985).

This pattern might be explained by a cost-of-adjustment model. If benefits from movement are linear and costs of adjustment increasing, and if the parameters are constant for the period under study, then each firm would adjust a fixed amount per period; that is, exhibit a local trend regardless of the behavior of other firms.

Our simple herd migration model offers an alternative explanation with additional linkages. Suppose, for the period studied, there is a linear per unit benefit from moving the debt-equity ratio toward its optimum (that is possibly firm-specific and time-varying), but a quadratic penalty for deviations from the crowd (i.e., other firms in one's industry). Under this scenario, parallel to the solution for linear-quadratic models for inventory, the firm's ideal ratio will be a linear weighting of its own past ratio plus the industry's expected ratio. We investigate the herd migration explanation by regressing the firm's debt-equity ratio on two own lags and one lag of the industry ratio. (The expected industry ratio is proxied by its lagged value; its contemporaneous value may exhibit a positive relation simply because of common shocks that influence the market value of equity across firms in an industry.)

A herd migration tendency is indicated by a significant positive sign (a *t*-statistic above +2) on the industry ratio. For 3 of the 10 industries, less than 15 percent of the firms exhibit such tendencies significantly. The proportions were significantly higher for the other 7 industries: food & tobacco, 11/17; paper, lumber & printing, 8/24; oil & gas, 7/13; chemicals, 7/24; electrical products & machinery, 7/12; transportation & communication, 7/17; wholesale & retail trade, 6/26. The median coefficient on the industry ratio for all 182 firms was 0.2; 35 percent of all the coefficients had *t*-statistics greater than +2, 23 percent were between +1 and +2, and 22 percent were negative.

IV. Concluding Remarks

Looking at flows in financial markets, a relatively unexplored area, we have illustrated the role of such behavioral phenomena as status quo bias, barn door closing, and herd migrations in influencing market outcomes. (These ideas will be elaborated in a forthcoming article in *Theory and Decision*.) The mere survival of many financial markets needs explanation. What sustains their flows of new funds? Absent substantial hedging activity, a large proportion of the individual participants (speculators) in a financial market must have negative expectations. Markets must have certain characteristics to continually lure in losing investors (Zeckhauser and V. Niederhoffer, 1983).

Monday morning quarterbacking may provide part of the explanation. Fans of professional football often believe (with the unacknowledged benefit of hindsight) that they would have been better able than their team's coaches/quarterbacks to identify the strategy that would have won the weekend game. In financial markets, similarly, participants examine past movements and convince themselves they would have made the right choices had they been involved, implying they could do so in the future. To confirm the prevalence of Monday morning quarterbacking, the reader should ask finance-oriented friends whether they sensed that the stock market was "clearly" overpriced just before the October 1987 crash.

We conclude poetically:

In the players, not the market, may rational ways inhere.

(But also vice versa—received doctrine makes it clear.)

From the traders to the tickers, you should not expect to see

Either easy aggregation, or pat synecdoche.

REFERENCES

- Camerer, C., "Do Biases in Probability Judgment Matter in Markets?," *American Economic Review*, December 1987, 77, 981-97.
- Kahneman, D., Slovic, P. and Tversky, A.,

- Judgement Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press, 1982.
- Patel, J., Zeckhauser, R. and Hendricks, D., "Investment Flows and Performance: Evidence from Mutual Funds, Cross-Border Investments, and New Issues," (1990) in a volume by the Center of Japan-U.S. Business and Economic Studies, New York University, forthcoming.
- Schelling, T., *Choice and Consequence*, Cambridge: Harvard University Press, 1982.
- Taggart, R., Jr., "Secular Patterns in the Financing of U.S. Corporations," in B. Friedman, ed., *Corporate Capital Structures in the United States*, Chicago: University of Chicago Press, 1985, 13-80.
- Zeckhauser, R. and Niederhoffer, V., "Futures Markets as Ecological Systems: Survival? Efficiency? Rational Participants?," presented at American Economic Association annual meeting, December 28, 1983.

Rational Addiction and the Effect of Price on Consumption

By GARY S. BECKER, MICHAEL GROSSMAN, AND KEVIN M. MURPHY*

Legalization of such substances as marijuana, heroin, and cocaine surely will reduce the prices of these harmful addictive drugs. By the law of the downward-sloping demand function, their consumption will rise. But by how much? According to conventional wisdom, the consumption of these illegal addictive substances is not responsive to price.

However, conventional wisdom is contradicted by Becker and Murphy's (1988) theoretical model of rational addiction. The Becker-Murphy (B-M) analysis implies that addictive substances are likely to be quite responsive to price. In this paper, we summarize B-M's model of rational addiction and the empirical evidence in support of it. We use the theory and evidence to draw highly tentative inferences concerning the effects of legalization of currently banned substances on consumption in the aggregate and for selected groups in the population.

Addictive behavior is usually assumed to involve both "reinforcement" and "tolerance." Reinforcement means that greater past consumption of addictive goods, such as drugs or cigarettes, increases the desire for present consumption. But tolerance cautions that the utility from a given amount of consumption is lower when past consumption is greater.

These aspects of addictive behavior imply several restrictions on the instantaneous utility function

$$(1) \quad U(t) = u[c(t), S(t), y(t)],$$

where $U(t)$ is utility at t , $c(t)$ is consumption of the addictive good, $y(t)$ is a non-addictive good, and $S(t)$ is the stock of "addictive capital" that depends on past consumption of c and on life cycle events. Tolerance is defined by $\partial u / \partial S = u_s < 0$, which means that addictions are harmful in the sense that greater past consumption of addictive goods lowers current utility. Stated differently, higher $c(t)$ lowers future utility by raising future values of S .

Reinforcement ($dc/dS > 0$) requires that an increase in past use raises the marginal utility of current consumption: ($\partial^2 u / \partial c \partial S = u_{cs} > 0$). This is a sufficient condition for myopic utility maximizers who do not consider the future consequences of their current behavior. But rational utility maximizers also consider the future harmful consequences of their current behavior. Reinforcement for them requires that the positive effect of an increase in $S(t)$ on the marginal utility of $c(t)$ exceeds the negative effect of higher $S(t)$ on the future harm from greater $c(t)$.

Becker-Murphy (p. 680) show that a necessary and sufficient condition for reinforcement near a steady state (where $c = \delta S$) is

$$(2) \quad (\sigma + 2\delta)u_{cs} > -u_{ss},$$

where u_{cs} and u_{ss} are local approximations near the steady state, σ is the rate of time preference, and δ is the rate of depreciation on addictive capital. Reinforcement is stronger, the bigger the left-hand side is relative to the right-hand side. Clearly,

[†]Discussants: Mark Kleiman, Harvard University; Robert Margo, Vanderbilt University; Peter Reuter, Rand Corporation.

*University of Chicago, Chicago, IL 60637, and NORC; City University of New York Graduate School and NBER; University of Chicago, NORC, and NBER, respectively. Our research has been supported by the Lynde and Harry Bradley Foundation through the Center for the Study of the Economy and the State, University of Chicago, and by the Hoover Institution.

$u_{cs} > 0$ is necessary if u is concave in S ($u_{ss} < 0$); that is, if tolerance increases as S increases.

It is not surprising that addiction is more likely for people who discount the future heavily (a higher σ) since they pay less attention to the adverse consequences. Addiction to a good is also stronger when the effects of past consumption depreciate more rapidly (δ is larger), for then current consumption has smaller negative effects on future utility. The harmful effects of smoking, drinking, and much drug use do generally disappear within a few years after a person stops the addiction unless vital organs, such as the liver, get irreversibly damaged.

Reinforcement as summarized in equation (2) has the important implication that the consumption of an addictive good at different times are complements. Therefore, an increase in either past or expected future prices decreases current consumption. The relation between these effects of past and future prices depends on both time preference and the depreciation rate.

Figure 1 illustrates several implications of our approach to addiction, where $S(t)$ is measured along the horizontal axis and $c(t)$ along the vertical one. The line $c = \delta S$ gives all possible steady states where c and S are constant over time. The positively sloped curves A^1 give the relation between c and S for an addicted consumer who has a particular utility function, faces given prices of c and y , and has a given wealth. The initial stock (S^0) depends on past consumption and past life cycle experience. Both c and S grow over time when S^0 is in the interval where A^1 is above the steady-state line, and both fall over time when S^0 is in the intervals where A^1 is below the steady-state line.

Figure 1 shows clearly why the degree of addiction is very sensitive to the initial level of addictive capital. If S^0 is below S^1 in the figure, a rational consumer eventually lays off the addictive good. But if S^0 is above S^1 , even a rational consumer becomes addicted, and ends up consuming large quantities of the addictive good.

The curve A^1 intersects the steady-state line at two points: $c^1 = \delta S^1$, and $c^{*1} = \delta S^{*1}$. Other relevant points are where $c = 0$ and

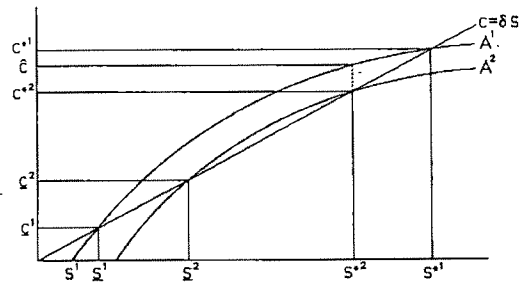


FIGURE 1

$S \leq S^1$. The second point and third set of points are locally stable. If initially $c = 0$, $S \leq S^1$, and a divorce or other events raise the stock of addictive capital to a level below S^1 , c may become positive, but eventually the consumer again refrains from consuming c . Similarly, if initially $c = c^{*1} = \delta S^{*1}$, c falls at first if say finding a good job lowers S from S^{*1} to a level $> S^1$. But c then begins to rise over time and returns toward c^{*1} . The other steady state, $c^1 = \delta S^1$, is locally and globally unstable: even small changes in S cause cumulative movements toward $c = 0$ or $c = c^{*1}$.

Unstable steady states are an important part of the analysis of rational addictions, for they explain why the same person is sometimes heavily addicted to cigarettes, drugs, or other goods, and yet at other times lays off completely. Suppose the consumer starts out at $c^{*1} = \delta S^{*1}$, and experiences favorable events that lower his stock of addictive capital below S^1 , the unstable steady state with A^1 . The consumer goes from being strongly addicted to eventually giving up c entirely. If A^1 is very steep when S is below the unstable steady state (if reinforcement is powerful in this interval), consumers would quit their addiction "cold turkey" (see the more extended analysis in B-M).

To analyze rational addicts' responses to changes in the cost of addictive goods, suppose they are at $c^{*2} = \delta S^{*2}$ along A^2 , and that a fall in the price of c raises the demand curve for c from A^2 to A^1 . Consumption increases at first from c^{*2} to \hat{c} , and then c grows further over time since \hat{c} is above the steady-state line. Consumption grows toward the new stable steady state at

$c^*1 = \delta S^*1$. This shows that long-run responses to price changes exceed short-run responses because initial increases in consumption of addictive goods cause a subsequent growth in the stocks of addictive capital, which then stimulates further growth in consumption.

Since the degree of addiction is stronger when A is steeper, and since long-run responses to price changes are also greater when A is steeper, strong addictions do not imply weak price elasticities. Indeed, if anything, rational addicts respond more to price changes in the long run than do nonaddicts.¹ The short-run change is smaller than the long-run change because the stock of addictive capital is fixed. Even in the short run, however, rational addicts respond to the anticipated growth in future consumption since future and current consumption of addictive goods are complements for them. But the *ratio* of short- to long-run responses does decline as the degree of addiction increases.²

The presence of unstable steady states for highly addictive goods means that the full effect of a price change on consumption could be much greater for these goods than the change between stable steady states given in footnote 1. Households with initial consumption capital between \underline{S}^2 and \underline{S}^1 in Figure 1 would be to the left of the unstable steady state at \underline{S}^2 when price equals p^2 , but they would be to the right of the unstable steady state at \underline{S}^1 when price equals p^1 . A reduction in price from p^2 to p^1 greatly raises the long-run demand by these households because they move from low initial

consumption to a stable steady state with a high level of consumption.

The total cost of addictive goods to consumers equals the sum of the good's price and the money value of any future adverse effects, such as the negative effects on earnings and health of smoking, heavy drinking, or dependence on crack. Either a higher price of the good (due perhaps to a larger tax) or a higher future cost (due perhaps to greater information about health hazards) reduces consumption in both the short and long run.

It is intuitively plausible that as price becomes a bigger share of total cost, long-run changes in demand induced by a given percentage change in the money price get larger *relative* to the long-run changes induced by an equal percentage change in future costs (see our 1991 paper, fn. 3). Money price tends to be relatively more important to poorer and younger consumers, partly because they generally place a smaller monetary value on health and other harmful future effects.

Poorer and younger persons also appear to discount the future more heavily (this is suggested by the theoretical analysis in Becker, 1990). It can be shown that addicts with higher discount rates respond more to changes in money prices of addictive goods, whereas addicts with lower rates of discount respond more to changes in the harmful future consequences.³

These implications of rational addiction can be tested with evidence on the demand for cigarettes, heavy consumption of alcohol, and gambling. Our earlier paper (1990)

¹Becker-Murphy show (equation (18), p. 685) that the long-run response between stable steady states to a permanent change in p_c is $dc^*/dp_c = \mu/\alpha_{cc}B'$, where μ is the marginal utility of wealth. The term B' measures the degree of addiction, where B' ranges between 1 (no addiction) and 0 for an addictive good that has a stable steady state.

²One can show that a rational addict's short-run response to a permanent change in p_c equals $dc_s/dp_c = -(\lambda/\delta)(dc^*/dp_c)$, where $-\delta \leq \lambda \leq 0$, and λ is larger when the degree of addiction is stronger (see B-M, pp. 679–80). Therefore, the ratio of the short- to long-term response gets smaller as the degree of addiction (measured by λ) is larger. But one can also show that dc_s/dp_c itself gets larger as the degree of addiction increases.

³If u is concave, $-\delta^2 u_{cc} - u_{ss} > 2\delta u_{cs}$. This implies that either or both of the following inequalities hold: $-u_{ss}/\delta^2 > u_{cs}/\delta$, and $-u_{cc} > u_{cs}/\delta$. We assume both hold. The second inequality states that an increase in c between steady states reduces the marginal utility of c by more than the increase in S raises it. The first inequality assumes that the increase in S has a larger effect on its marginal utility than does the increase in c .

The absolute value of the long-run change in c induced by a change in p_c is raised by an increase in σ if $-u_{ss} > \delta u_{cs}$. Similarly, the absolute value of the long-run change in c with respect a change in future costs is reduced by an increase in σ if $-u_{cc}\delta > u_{cs}$. (For more details, see our 1991 paper, fn. 4).

fit models of rational addiction to cigarettes to a time-series of state cross sections for the period 1955–85. We find a sizable long-run price elasticity of demand ranging between $-.7$ and $-.8$, while the elasticity of consumption with respect to price in the first year after a permanent price change (the short-run price elasticity) is about $-.4$. Smoking in different years appear to be complements: cigarette consumption in any year is lower when both future prices and past prices are higher.

Frank Chaloupka (forthcoming) analyzes cigarette smoking over time by a panel of individuals. He finds similar short- and long-run price elasticities to those we estimate, and that future as well as past increases in cigarette prices reduce current smoking. He also finds that smoking by the less educated responds much more to changes in cigarette prices than does smoking by the more educated; a similar result has been obtained by Joy Townsend (1987) with British data. Eugene Lewit et al. (1981) and Lewit and Douglas Coate (1982) report that youths respond more than adults to changes in cigarette prices. By contrast, the information that began to emerge in the early 1960's about the harmful long-run effects of smoking has had a much greater effect on smoking by the rich and more educated than by the poor and less educated (see Phillip Farrell and Victor Fuchs, 1982, for the United States; Townsend for Britain).

Philip Cook and George Tauchen (1982) examine variations in death rates from cirrhosis of the liver (a standard measure of heavy alcohol use), as well as variations in per capita consumption of distilled spirits in a time-series of state cross sections for 1962–77. They find that state excise taxes on distilled spirits have a negative and statistically significant effect on the cirrhosis death rate. Moreover, a small increase in prices in a state's excise tax lowers death rates by a larger percentage than it lowers per capita consumption.

Pamela Mobilia (1990) applies the rational addiction framework to the demand for gambling at horse racing tracks. Her data consist of a U.S. time-series of racing track cross sections for the period 1950–86 (tracks

over time are the units of observation). She measures consumption by the real amount bet per person attending (handle per attendant), and price by the takeout rate (the fraction of the total amount bet that is retained by the track). Her findings are similar to those in the rational addictive studies of cigarettes. The long-run price elasticity of demand for gambling equals $-.7$ and is more than twice as large as the short-run elasticity of $-.3$. Moreover, an increase in the current takeout rate lowers handle per attendant in both past and future years.

The evidence from smoking, heavy drinking, and gambling rather strongly supports our model of rational addiction. In particular, long-run price elasticities are sizable and much bigger than short-run elasticities, higher future as well as past prices reduce current consumption, lower-income persons respond more to changes in prices of addictive goods than do higher-income persons, whereas the latter respond more to changes in future harmful effects, and younger persons respond more to price changes than older persons. It seems reasonable to us that what holds for smoking, heavy drinking, and gambling tends to hold also for drug use, although direct evidence is not yet available, and many experts on drugs would be skeptical. Lacking the evidence, we simply indicate what to expect from various kinds of price changes if responses of drug addicts are similar to those of persons addicted to other goods.

To fix ideas, consider a large permanent reduction in the price of drugs (perhaps due to partial or complete legalization) combined with much greater efforts to educate the population about the harm from drug use. Our analysis predicts that much lower prices could significantly expand use even in the short run, and it would surely stimulate much greater addiction in the long run. Note, however, that the elasticity of response to large price changes would be less than that to modest changes if the elasticity is smaller at lower prices.

The effects of a fall in drug prices on demand would be countered by the education program. But since drug use by the poor would be more sensitive to the price fall than to greater information about harm-

ful longer-run effects, drug addiction among the poor is likely to become more important relative to addiction among the middle classes and rich. For similar reasons, addiction among the young may rise more than that among other segments of the population.

A misleading impression about the reaction to permanent price changes may have been created by the effects of temporary police crackdowns on drugs, or temporary federal "wars" on drugs. Since temporary policies raise current but not future prices (they would even lower future prices if drug inventories are built up during a crackdown period), there is no complementary fall in current use from a fall in future use. Consequently, even if drug addicts are rational, a temporary war that greatly raised street prices of drugs may well have only a small effect on drug use, whereas a permanent war could have much bigger effects, even in the short run.

Clearly, we have not provided enough evidence to evaluate whether or not the use of heroin, cocaine, and other drugs should be legalized. A cost-benefit analysis of many effects is needed to decide between a regime in which drugs are legal and one in which they are not. What this paper shows is that the permanent reduction in price caused by legalization is likely to have a substantial positive effect on use, particularly among the poor and young.

REFERENCES

- Becker, Gary S., "Optimal Discounting of the Future," Department of Economics, University of Chicago, April 1990.
- _____, Grossman, Michael and Murphy, Kevin M., "An Empirical Analysis of Cigarette Addiction," NBER Working Paper No. 3322, April 1990.
- _____, _____, and _____, "Rational Addiction and the Effect of Price on Consumption," Working Paper, Center for the Study of the Economy and the State, University of Chicago, 1991.
- _____, and Murphy, Kevin M., "A Theory of Rational Addiction," *Journal of Political Economy*, August 1988, 96, 675-700.
- Chaloupka, Frank J., "Rational Addictive Behavior and Cigarette Smoking," *Journal of Political Economy*, forthcoming.
- Cook, Philip J. and Tauchen, George, "The Effect of Liquor Taxes on Heavy Drinking," *Bell Journal of Economics*, Autumn 1982, 13, 379-90.
- Farrell, Phillip and Fuchs, Victor R., "Schooling and Health: The Cigarette Connection," *Journal of Health Economics* December 1982, 1, 217-30.
- Lewit, Eugene M. and Coate, Douglas, "The Potential for Using Excise Taxes to Reduce Smoking," *Journal of Health Economics* August 1982, 1, 121-45.
- _____, _____, and Grossman, Michael, "The Effects of Government Regulation on Teenage Smoking," *Journal of Law and Economics*, December 1981, 24, 545-69.
- Mobilia, Pamela, "An Economic Analysis of Addictive Behavior: The Case of Gambling," unpublished doctoral dissertation, City University of New York, 1990.
- Townsend, Joy L., "Cigarette Tax, Economic Welfare and Social Class Patterns of Smoking," *Applied Economics*, 1987, 19, 355-365.

Alcohol Consumption During Prohibition

By JEFFREY A. MIRON AND JEFFREY ZWIEBEL*

The burgeoning debate over drug legalization in the United States has drawn renewed attention to the nation's experience with Prohibition. Although the parallels between the criminalization of alcohol and the criminalization of drugs are not exact, Prohibition provides a natural setting in which to examine the impact of legal restrictions on the use of substances such as alcohol or drugs. The popular media asserts widely divergent accounts of the changes in alcohol consumption during Prohibition, claiming both that drinking increased substantially and that drinking fell to a small fraction of its pre-Prohibition level. To date, however, most such assertions have been based on little hard evidence.

It should come as no surprise that accurate data on alcohol consumption during Prohibition do not exist. Perhaps more surprisingly, there have been few serious attempts to estimate consumption using related statistics. With the notable exception of Clark Warburton (1932), which has the drawback of being conducted in the middle of Prohibition, we know of no careful attempt to estimate this consumption. We employ Warburton both as a starting point and as a comparison for our estimation.

Attempts to estimate alcohol consumption from related variables suffer the drawback that Prohibition may have altered the relationship between these series and alcohol consumption. We address this problem by using data drawn from widely varying sources; plausibly the biases in these series will be unrelated. In particular, we use mortality, mental health, and crime statistics to

estimate the consumption of alcohol during Prohibition.

We find that alcohol consumption fell sharply at the beginning of Prohibition, to approximately 30 percent of its pre-Prohibition level. During the next several years, however, alcohol consumption increased sharply, to about 60–70 percent of its pre-Prohibition level. The level of consumption remained virtually the same immediately after Prohibition as during the latter part of Prohibition, although consumption increased to approximately its pre-Prohibition level during the subsequent decade.

I. Historical Background

The Prohibition movement in the United States traces its origins to the mid-nineteenth century. It was not until the 1910's, however, that sufficient support was garnered to make national prohibition a reality. During the latter half of this decade, many states enacted dry laws, and in 1917 Congress provided for Wartime Prohibition. National Prohibition became effective in January 1920 under the 18th Amendment to the Constitution. Prohibition remained in effect for almost 14 years, until rescinded by the 21st Amendment in December 1933.

By the mid-1920's it was apparent that at best limited success had been achieved in prohibiting alcohol consumption. Initially Congress responded with increased enforcement. Money appropriated for enforcing Prohibition increased from \$6.3 million in 1921 (the first year of large-scale enforcement) to \$9.2 million in 1925 and to \$13.4 million in 1930 (U.S. Department of Treasury, 1930, p. 2). However, the inability to restrict the illegal trade and the inevitable accompanying corruption eventually led to widespread public disenchantment with Prohibition.

By the turn of the decade, popular sentiment had undergone a radical turnabout on

*Department of Economics, Boston University, Boston, MA 02215 and Department of Economics, MIT, Cambridge, MA 02139, respectively. Zwiebel acknowledges financial support from the National Science Foundation and the Alfred P. Sloan Foundation. We thank Peter Temin and Robert Margo for helpful comments.

Prohibition.¹ The 1930 election saw the anti-Prohibitionists' strength increase, and by 1932 the Democratic Party supported outright repeal. By 1933, support for repeal was widespread in Congress. In February, both Houses approved the 21st Amendment, and by December, three-quarters of the states had ratified the amendment, ending the experiment of Prohibition.

II. Data and Methodology

Estimating alcohol consumption during Prohibition is complicated by the possibility that Prohibition was accompanied by changes in attitudes or actions that affected underlying relationships. Thus, for example, while the number of arrests for drunkenness may be closely related to alcohol consumption, Prohibition could lead to more vigorous enforcement of drunkenness laws, raising the number of drunkenness arrests for a fixed level of consumption. Alternatively, it could drive more drinking into the home, thereby lowering the drunkenness arrest tally. Similarly, deaths due to alcoholism may increase due to low-quality alcohol. We address such complications by comparing estimates from several diverse sources.

The series that we use to estimate alcohol consumption are the death rate from cirrhosis of the liver, the death rate from alcoholism, the number of patients per capita admitted to hospitals for the first time with alcoholic psychosis, and the rate of drunkenness arrests. (The Data Appendix, available upon request, provides the details of the construction of these series.) For each of the series, we posit the relation.

$$(1) \quad \ln Y_t = \alpha + \beta t + \gamma \ln X_t + \varepsilon_t,$$

¹For example, in 1915 popular magazine articles in favor of Prohibition outnumbered those opposed 20 to 1; by 1930 this ratio had reversed to 1 to 2 (see Andrew Sinclair, 1962, p. 332.). Polls taken by *Literary Digest* indicate that while in 1922 only 1 in 5 individuals favored complete repeal, by 1930 all states but 5 showed a majority in favor of repeal or modification, and by 1932 all states but 2 had a majority in favor of repeal (Sinclair, p. 335).

where X_t is alcohol consumption, t is a time trend, and Y_t is one of the four series related to alcohol consumption. We estimate this equation for the years during the 1900–50 period for which data for the particular series are available, exclusive of the Prohibition years 1920–35.² We then use the estimated parameters from (1) to construct consumption from 1920 to 1935.

While Warburton assumes linear relationships between alcohol consumption and various statistics, we assume a log linear relationship because this is the simplest specification satisfying the restriction that no alcohol consumption should imply no deaths from alcoholism, no alcoholic psychosis and no drunkenness arrests (although the same cannot be said about cirrhosis). Additionally, this model fits well for all four series, and a linear specification yields similar results. The trend is included to capture other developments over time, such as a tendency to treat more psychotic patients or an improvement in the treatment of cirrhosis.

Including lagged consumption in these regressions does not substantively change the results. For cirrhosis and drunkenness, the coefficients on lagged consumption are statistically insignificant. For deaths due to alcoholism and admittances for alcoholic psychosis, lagged values of alcohol consumption do enter significantly, but the estimates of alcohol consumption based on regressions with these lags are not substantially different from those based on the regressions without lags.

The fact that lagged consumption does not explain cirrhosis may appear surprising, since cirrhosis results from a lengthy history of alcohol consumption. While this may be

²The estimates reported below are robust to extending the sample period. We use data only through 1950 to minimize the effects of changes in underlying relationships and to avoid definitional changes in the reported data. We exclude 1934–35 from the sample because it took several years after the end of Prohibition for the legal producers to fully recapture industry control. For alcoholic psychosis, we fit the model only through 1940 because this series (like other mental health series) is quite volatile during World War II.

so, the data seem to suggest that one must be presently drinking to die from cirrhosis. This view is mirrored in statistics that show a steep drop in cirrhosis when consumption falls both during wartime Prohibition and at the onset of constitutional Prohibition. If, however, the true specification for cirrhosis involves lags that we do not include, our estimates of consumption are likely overstated immediately after the onset of Prohibition (when consumption falls) and understated in following years.³

III. Results

Table 1 presents ordinary least squares estimates of equation (1). Each row represents a regression with one of the four dependent variables. Low Durbin-Watson statistics indicate the likelihood of serial correlation, so we report robust standard errors calculated using Whitney Newey and Kenneth West's (1987) procedure. For each of the four series, the model explains a large portion of the variation in the dependent variable. The \bar{R}^2 exceeds .90 for alcoholism, cirrhosis, and drunkenness and is .79 for alcoholic psychosis. Alcohol consumption is significant at the .01 level of significance for all series except alcoholism deaths, for which it is significant at the .03 level. The time trend is significant at the .01 level for cirrhosis, alcoholism, and psychosis, but is insignificant at the .05 level for drunkenness arrests.

The four estimates of consumption from 1920 to 1935, and true consumption before and after this period, are graphed in Figure 1. Comparing the four series we find similar estimates from cirrhosis, drunkenness, and psychosis, but substantially higher estimates from alcoholism. We suspect that the alcoholism series overstates true consumption during Prohibition due to decreased alcohol quality. In particular, the consumption of wood or denatured alcohol likely produced

TABLE 1—ESTIMATES OF EQUATION (1)

Dependent Variable	Sample Period	Independent Variables			\bar{R}^2
		Constant	Trend	Alcohol	
Cirrhosis	1900–50	2.560 (.046)	–.007 (.001)	.619 (.046)	.924
Alcoholism Deaths	1900–50	2.366 (.221)	–.027 (.002)	.802 (.344)	.902
Drunkenness Arrests	1910–29	4.186 (.345)	.013 (.007)	.902 (.076)	.933
Alcoholic Psychosis	1910–40	.691 (.123)	.008 (.002)	.949 (.078)	.794

Notes: 1) Newey-West standard errors are shown in parentheses. 2) Alcohol consumption is measured in gallons of pure alcohol per capita. 3) Cirrhosis, alcoholism deaths, alcoholic psychosis, and drunkenness arrests are all measured in per capita terms. 4) The equations are estimated over the sample periods indicated, excluding the years 1920–35.

more alcoholism deaths for given consumption. Similarly, cirrhosis, which provides the lowest estimates, may understate consumption if the functional relationship is misspecified, as discussed above.

All four estimates, however, show a similar steep initial decline in consumption followed by a steady increase. Consumption falls immediately after enactment of Prohibition to 20 to 40 percent of its pre-Prohibition level. Alcoholism, drunkenness, and psychosis estimates indicate a sharp rebound in consumption from 1921 to 1927 and a less dramatic increase after 1927. The cirrhosis estimates exhibit a similar pattern, but with a smaller initial decline in consumption and a more moderate subsequent increase. In the later years of Prohibition, cirrhosis, drunkenness, and psychosis estimate consumption to be 50 to 70 percent of its pre-Prohibition value, while alcoholism estimates small increases in consumption.

The estimates in Figure 1 improve on Warburton's by employing data beyond 1929 (both in fitting the model and in estimating consumption at the end of Prohibition) and by considering a more reasonable functional relationship. Nonetheless, both studies yield

³For further discussion of this issue and a more detailed model of the relation between alcohol consumption and the cirrhosis death rate, see P. J. Cook and G. Tauchen (1982).

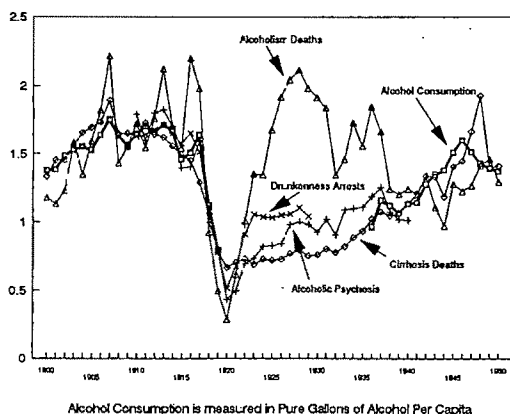


FIGURE 1. ESTIMATED ALCOHOL CONSUMPTION

similar results. Warburton considers agricultural sources of production, death rates, and arrests for drunkenness in estimating consumption. He estimates that consumption per capita is around 65 percent of pre-Prohibition levels by 1925 and around 71 percent by 1929. When comparing results from the same series, our estimates are slightly higher than Warburton's. Overall, however, his average estimates are about the same as ours because his highest estimates are from agricultural production, which we do not consider.

While for three of our series we find reductions in consumption compared to pre-Prohibition levels, the decline is much more modest when compared to post-Prohibition levels. The level of consumption in 1937–40 is about the same as our average estimate for the last years of Prohibition. (However, consumption rises to pre-Prohibition levels over the next decade.) Whether the pre- or post-Prohibition benchmark is appropriate depends both on what question is being asked and to what one attributes the difference in pre- vs. post-Prohibition consumption. This difference may result from demographic factors (for example, a smaller percentage of immigrants who drank more, or a different age composition of the population), or a continuation of the social trend toward less drinking that began well before Prohibition. Either explanation would imply that Prohi-

bition had little to do with the observed change in drinking patterns. Conversely, the difference may result from a change in social attitudes due to Prohibition. Trying to distinguish between these competing hypotheses is beyond the scope of this paper. However, as far as the debate on drug legalization is concerned, the comparisons to post-Prohibition consumption are more pertinent than those to pre-Prohibition consumption.

IV. Discussion

There are several channels through which Prohibition may affect alcohol consumption. First, Prohibition increases supply costs, as these must include the cost of evading detection and the potential cost of punishment. This implies a higher equilibrium market price and less consumption. Second, Prohibition inhibits consumer access to alcohol by raising search costs, making quality dubious, and increasing the possibility of being cheated. Third, Prohibition may create a prevailing sentiment that a certain good is "bad" or "immoral," thereby decreasing consumer demand. Finally, Prohibition may deter some individuals' consumption because of "respect for the law." Even though consumption per se was not illegal, purchasing alcohol during Prohibition involved doing business with criminals.

Our results suggest that of these reasons, only the first two contributed significantly to the changes in alcohol consumption during Prohibition. A careful consideration of price quotes in newspapers by Warburton suggests that prices in 1930 were approximately three times as high as pre-Prohibition prices.⁴ Hence even if price changes alone were responsible for changes in demand,

⁴See Warburton (pp. 113, 116, and 166). We assume that the cost of homemade alcohol was at least as high as the market price after accounting for time and potential punishment costs. If it had been much cheaper, there would not have been an illegal alcohol industry. Note that another interpretation of higher prices and lower consumption during Prohibition is that illegal suppliers possessed and exercised monopoly power.

global price elasticities would have to be extremely low, around .1.⁵ This suggests that the effect of all other avenues that could theoretically lower demand had a negligible impact.

This is consistent with anecdotal evidence that suggests that the effect of public sentiment in reducing consumption is unclear. Some evidence even suggests Prohibition made consumption more desirable by endowing drinking with an illicit romance and sense of adventure. Thus, one plausible interpretation of the small changes in consumption given the change in price is that the demand curve for alcohol shifted out during Prohibition.

There are important similarities and differences to keep in mind when trying to draw inferences from Prohibition on how drug legalization might change consumption. Prices of illegal drugs appear to have been forced further above their production costs than that of alcohol during Prohibition, presumably because of more stringent enforcement. This effect, however, may be countered by a more inelastic demand for illegal drugs than for alcohol. There seems to be no compelling reason why respect for the law or other social impediments are any more likely to have a significant impact on drug consumption than they did on alcohol consumption during Prohibition. Thus, we hypothesize that any increase in consumption due to changes in social attitudes following drug legalization is likely to be small.

V. Conclusion

We find that while alcohol consumption declined sharply at the onset of Prohibition,

within several years it rebounded to 60–70 percent of its initial value and did not increase substantially immediately following the repeal of Prohibition. Claims either that consumption during Prohibition increased significantly or that it fell to a small fraction of previous usage can be patently rejected. Changes in consumption during Prohibition were modest given the change in price. This suggests that legal deterrents had little effect on limiting consumption outside of their effect on price. Social pressure and respect for the law did not go far in reducing consumption during Prohibition. We speculate that this is likely to be true as well with illegal drugs today, and therefore claims based on such arguments exaggerate the extent to which drug consumption would increase upon legalization.

Of course, any debate on drug legalization is incomplete if it solely considers changes in consumption. The negative effects accompanying any increases in consumption are costs that have to be weighed against various benefits of drug legalization. These benefits are likely to include an elimination of the violent drug culture that results from the battle for illegal profits, a reduction in overdoses from impure drugs, a reduction in robberies and burglaries committed by addicts who pay inflated drug prices, the stabilization of Latin American regimes fighting control battles with drug lords, the ability to combat the spread of AIDS from needle exchanges more effectively, and an unclogging of the criminal justice system. This paper does not attempt to calculate the costs and benefits of legalization. Rather, it suggests that if Prohibition is any guide, the cost to society from increased drug use is likely to be smaller than commonly believed.

REFERENCES

- Cook, P. J. and Tauchen, G., "The Effect of Liquor taxes on Heavy Drinking," *Bell Journal of Economics*, Autumn 1982, 13, 379–390.
- Newey, Whitney and West, Kenneth, "A Simple, Positive Definite, Heteroskedasticity and

⁵Recent estimates of the demand elasticity for alcohol vary greatly. For example, Cook and Tauchen estimate an elasticity of 1.8, while S. I. Ornstein and D. M. Hanssens (1985) estimate elasticities of .8 to 1.0 for spirits but only .1 for beer. In any event, the applicability of these figures here is questionable. These studies estimate local elasticities by considering tax changes. However, alcohol elasticities are unlikely to be constant over a wide range of prices and may have changed significantly over time. Furthermore, the above cited studies, in addition to having conflicting results, are plagued by the lack of reliable price data.

- Autocorrelation Consistent Covariance Matrix," *Econometrica*, May 1987, 55, 703-08.
- Ornstein, S. I. and Hanssens, D. M., "Alcohol Control Laws and the Consumption of Distilled Spirits and Beer," *Journal of Consumer Research*, September 1985, 12, 200-13.
- Sinclair, Andrew, *Prohibition the Era of Excess*, London: Faber and Faber, 1962.
- Warburton, Clark, *The Economic Results of Prohibition*, New York: Columbia University Press, 1932.
- U.S. Department of Treasury, Statistics Concerning Intoxicating Liquors, Washington: USGPO, 1930.

Who Uses Illegal Drugs?

By ROBIN SICKLES AND PAUL TAUBMAN*

In this paper we will use the *National Longitudinal Sample of Youth* (NLSY) to estimate a model of who uses various drugs and at what ages. Data already exist on who takes illegal drugs, but thus far the descriptive analysis has not utilized recently developed statistical frameworks that provide for more informative estimates.

I. Prior Literature

Gary Becker and Kevin Murphy (1988) present a theoretical model of rational addiction that requires information on past, present and future prices. They do not study illegal drugs empirically in this and their related papers.

A major data source is the annual survey of high school seniors (HSS), also known as "Monitoring the Future." This work is summarized in L. Johnston et al. (1988). They find a growing use of drugs (measured by monthly, annual, and lifetime prevalence) over time, and differences by region and sex. Cocaine use showed marked increases from 1976, though this levelled off from 1986 to 1987. They often rely on cross-tabs that leaves many variables uncontrolled and the results subject to omitted variable bias.

J. Bachman et al. (1984) use ordinary least squares (OLS) regressions in which the drug use in the three years post-high school is related to various characteristics such as living arrangements. While an improvement over cross-tabs, OLS applied to categorical dependent variables yields inefficient estimates (see M. Nerlove and S. Press, 1973).

Johnston et al. used follow-up surveys of a subsample drawn from each cohort. They find the use of some drugs decline at older

ages, say 35, though they do not determine if the heavy users have died, dropped out of the sample, or have been rehabilitated.¹

The HSS's initial restriction to high school seniors removes about 30 percent of the population who drop out of high school perhaps because of taking drugs. Studies based on the National Institute of Drug Abuse (NIDA) sample of people 12 and older show some heavy drug use of people less than 17 years old. (See J. D. Miller et al., 1982, and NIDA, 1985.)

Richard Clayton (1985) using the 1980 HSS presents univariate regressions that shows the frequency of use of cocaine is positively related to lifetime marijuana use and days out of school in the past month, but negatively related to high school grade point average. H. Abelson and Miller (1985), using the NIDA surveys covering 1974–82, show differences in the percentage using cocaine by education and race—for lifetime, 12 months, and last month measures. They find strong trends.

D. Kandel (1980) presents a recent survey of drinking and drug use among youth. People in their late 30's "mature out" of heroin use rather than die.

II. The Model

We assume that individuals maximize a utility function that has, as its arguments, various types of consumption and leisure. Note that some of the C_i s may affect utility by their impact on health in the utility function.

The choices of the level of the various consumption goods and leisure are subject to a budget constraint and a health production function. The health production function relates the stock or level of health to

*Rice University, Houston, TX 77251 and NBER, and University of Pennsylvania, Philadelphia, PA 19104 and NBER, respectively. We thank Craig Strain for his valuable research assistance.

¹The rehabilitation literature as in J. F. Maddux and D. P. Desmond (1986) is not very encouraging on this score.

a person's endowments (E) and some of the C_i .

The constrained maximization of utility leads to demand functions for each of the C_i including drugs which will depend on the wage rate, endowments, prices, and tastes. We allow tastes to vary by socioeconomic status such as education, family background, marital status, and age. Family background measures will include parental education and religious preference. We do not include recently obtained prices at this stage of our study and thus the interpretation of our results as indicative of demand behaviors rests on the strong assumptions of highly price inelastic demand behavior.

The National Longitudinal Survey (NLS) contains several random samples of men and women in various age groups. Data on drug use have been collected in the youth segment, which started in 1979, in two time periods—1984 and 1988. Date of first use of various drugs is also available. For each sample, interviews have been conducted over time on many variables other than drugs.

This sample has drug use information measured over the last month, last year, and lifetime, and also amount used for a number of specific drugs including amphetamines, barbituates, heroin, and cocaine. We study separately marijuana, a combination of marijuana, cocaine, and all other drugs (including about 55 cases of barbituate use), and no drug use during the last 12 months. (Most of the hard drug observations are for cocaine.)

Self-Reported Drug Data Accuracy. Our sample uses surveys in which people are asked if they have taken various illegal drugs. Will such questions elicit valid and reliable information? Zili Amsel et al. (1976) concluded that such self-reported drug use data were fairly accurate using a sample of 829 addicts under treatment.

Johnston et al. used several waves of responses to the drug use questions in "Monitoring The Future," and conclude that either people lie consistently over time or provide reasonably accurate answers. They state: "Like most studies dealing with sensitive behaviors, we have no direct, objective

validation of the present measures; however, the considerable amount of inferential evidence that exists strongly suggests that the self-report questions produce largely valid data" (p. 20). Moreover, a majority of seniors, and up to 80 percent of the subsequent subsample, report taking drugs.

E. Wish (1987) studied men recently arrested and held in the Manhattan Central Booking in 1984. About 95 percent of those approached agreed to an interview and 84 percent of these gave a urine specimen. The interview data yielded substantially smaller drug use than the urine specimens (for cocaine in 1986, 43 vs. 82 percent), but as Wish notes, the circumstances are not conducive to honesty.

B. S. Mensch and D. B. Kandel (1988) argue that *NLS* drug use reports are too low because of shame associated with admitting to partaking in an immoral activity to an interviewer whom you see annually. However, the head of the *NLS*, R. Olsen, has informed us of an unpublished study in which half the respondents were asked the questions by the interviewer and half were given a questionnaire to mail back. The drug use estimates were the same via the two methods, but the mailed-back questionnaire reported significantly more abortions.

Statistical Methods. Prior research has generally used estimates based on means for various groups or on OLS regressions. Using means for various groups is only useful if all cells have large numbers of observations, but available data sets are too small to allow for more than a few characteristics in defining a cell. The linear probability model has a number of statistical problems (G. S. Maddala, 1983).

We allow the demand response, Y , of an individual unit to be restricted to one of a number, say $k + 1$ ($k \geq 1$), of ordinal values, denoted for convenience by $1, \dots, k, k + 1$. Using the LOGISTIC model in "SAS," we fit a parallel lines regression model based on the cumulative distribution probabilities of the response categories, rather than on their individual probabilities. The model has the form

$$g(\Pr(Y \leq i|x)) = \alpha_i + \beta'x, \quad 1 \leq i \leq k$$

TABLE 1—ORDERED LOGITS FOR 1984

Variable	Analysis of Maximum Likelihood Estimates			
	Coeff. Est.	Std. Error	Wald Chi-Sq.	Pr > Chi-Sq.
INTERCP1	2.697	0.528	26.134	0.0001
INTERCP2	4.330	0.529	66.986	0.0001
AGE	-.07	0.011	43.338	0.0001
MGRADE	-0.054	0.0097	31.342	0.0001
FGRADE	-0.053	0.0075	49.289	0.0001
FE	-0.012	.0024	24.835	0.0001
BLACK	0.349	.06	34.189	0.0001
HISP	.279	0.079	12.328	0.0004
INC84	6.95E-7	3.75E-6	0.034	0.8530
ED	-0.403	.076	28.317	0.0001
EDSQ	0.018	0.003	36.685	0.0001
SEX	.313	.045	48.689	0.0001
CATH	-0.189	.057	11.169	0.0008
NORL	-0.653	0.078	69.929	0.0001
BABT	0.075	.061	1.525	0.2168

Association of Predicted Probabilities and Observed Responses:

Concordant = 62.4%	Somers D = 0.252
Discordant = 37.2%	Gamma = 0.253
Tied = 0.5%	Tau-a = 0.626
	c = 0.626

TABLE 2—ORDERED LOGITS FOR 1988

Variable	Analysis of Maximum Likelihood Estimates			
	Coeff. Est.	Std. Error	Wald Chi-Sq.	Pr > Chi-Sq.
INTERCP1	-0.096	0.493	.038	0.8461
INTERCP2	1.360	0.493	7.599	0.0058
AGE	-0.0042	0.0096	0.195	0.6591
MGRADE	-0.055	0.0096	32.749	0.0001
FGRADE	-0.038	0.0075	25.853	0.0001
FE	-0.0042	0.0023	3.385	0.0658
BLACK	0.125	0.059	4.419	0.0355
HISP	-.012	0.079	.022	0.8834
INC84	-4.4E-6	2.4E-6	3.565	0.0590
ED	-0.048	.066	.528	0.4677
EDSQ	.0032	0.0024	1.676	0.1955
SEX	0.263	0.045	33.606	0.0001
CATH	-0.205	.057	13.107	0.0003
NORL	-.384	.078	24.481	0.0001
BABT	.109	.061	3.179	0.0746

Association of Predicted Probabilities and Observed Responses:

Concordant = 59.1%	Somers D = 0.188
Discordant = 40.3%	Gamma = 0.189
Tied = 0.7%	Tau-a = 0.125
	c = 0.626

where $\alpha_1, \dots, \alpha_k$ are k intercept parameters, and β is the vector of slope parameters.

The logit function $g(\rho) = \log(\rho/(1-\rho))$ is the inverse of the cumulative logistic distribution function, which is $F(x) = 1/(1 + \exp(-x))$.

Suppose the response variable can take on the ordered values $1, \dots, k, k+1$. The probability that the j th observation has response i is given by

$$\Pr(Y_j = i | x_j) = \begin{cases} F(\alpha_1 + \beta'x_j) & i = 1 \\ F(\alpha_i + \beta'x_j) - F(\alpha_{i-1} + \beta'x_j) & 1 < i \leq k \\ 1 - F(\alpha_k + \beta'x_j) & i = k+1 \end{cases}$$

In our estimates we have included an estimate of each individuals' fixed effect obtained from a Tobit estimate of a wage equation estimated over the period 1979-83. Here we use the independent variables in Tables 1 and 2 and family income. (Results from other specifications that alter the num-

ber of demand categories and the distribution of latent demand are available on request.) We average the residuals over time for each individual. (We used the inverse of the Mills ratio to correct for selectivity.) This estimated fixed effect is denoted as FE .

III. Results

We have calculated means and standard deviations in 1984 and 1988. We code drug use as in the last year: no use (about 30 percent in both years), only marijuana (about 35 and 30 percent in 1984 and 1988, respectively), and marijuana and other drugs, mostly cocaine. (Means for other variables are given in the Appendix, available on request.)

Tables 1 and 2 present our main findings based on ordered logits (ranging from no use = 0 to marijuana plus cocaine = 2). *MGRADE* and *FGRADE* are years of education of the mother and father and the religious preference dummies are Catholic, no religion, and Baptist. Other variable names are obvious. In 1984, all variables are statistically significant except income in

1984 and belonging to the baptist religion. Taking account of the ordering scale, use of drugs decreases with education of self and parents, wage fixed effects, and being Catholic or no religious preference. There are increases with being black or female.

In 1988 the sign and significance patterns are the same though AGE is now of marginal significance. The 1988 coefficients are generally of smaller size in absolute value.

The relative probability of moving from no drug use to moderate drug use in 1984 is equal to 11.6 times the coefficients shown in Table 1. For example, blacks are 400 percent more likely to make this move. The movement to hard drugs from the no use category is 3.2 times the coefficient or about 120 percent for blacks. (Calculated at the means for all other variables.) There are both bigger and smaller effects for other variables.

In both years, we have examined the results on other coefficients of dropping the wage fixed effects variable. Some of the coefficients such as black and income change greatly.

There are many significant effects of sociodemographic variables on drug use. The results seem stable over our two time periods. Estimates of income fixed effects are quite important with people who earn more using drugs more, though the wage fixed effects are weaker in 1988.

REFERENCES

- Abelson, H. and Miller, J. D., "A Decade of Trends in Cocaine Use in the Household Population," in E. Adams and N. Kozel, eds., *Cocaine Use in America: Epidemiologic and Clinical Perspectives*, Rockville: NIDA, Research Monograph Series 61, 1985.
- Amsel, Zili et al., "Reliability and Validity of Self-Reported Illegal Activities and Drug Use Collected from Narcotic Addicts," *International Journal of the Addictions*, March/April 1976, 11, 325-36.
- Bachman, J., O'Malley, P. and Johnston, L., "Drug Use Among Young Adults: The Impact of Role Status and Social Environment," *Journal of Personality and Social Psychology*, March 1984, 47, 629-45.
- Becker, Gary and Murphy, Kevin M., "A Theory of Rational Addiction," *Journal of Political Economy*, August 1988, 96, 675-700.
- Clayton, Richard R., "Cocaine Use in the United States: In a Blizzard or Just Being Snowed?," in *Cocaine Use in America: Epidemiologic and Clinical Perspectives*, Rockville: NIDA, Research Monograph Series 61, 1985.
- Johnston, L., O'Malley, P. and Bachman, J., *Illicit Drug Use, Smoking and Drinking By America's High School Students, College Students, and Young Adults*, Washington: U.S. Department of Health and Human Services, 1988.
- Kandel, D., "Drug and Drinking Behavior Among Youth," *Annual Review of Sociology*, 1980, 6, 235-83.
- Maddala, G. S., *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press, 1983.
- Maddux, J. F. and Desmond, D. P., "Relapse and Recovery in Substance Abuse Careers," in F. M. Tims and C. G. Leukefeld, eds., *Relapse and Recovery in Drug Abuse*, Washington: Division of Clinical Research, NIDA Research Monograph 72, 1986.
- Mensch, B. S. and Kandel, D. B., "Do Job Conditions Influence the Use of Drugs?," *Journal of Health and Social Behavior*, June 1988, 29, 169-84.
- Miller, J. D. et al., *National Survey on Drug Abuse: Main Findings*, Rockville: NIDA, 1982.
- Nerlove, M. and Press, S., *Univariate and Multivariate Log-Linear and Logistic Models*, Rand R1306-EPA/NIH, 1973.
- Wish, E., *Drug Use Forecasting: New York 1984 to 1986*, Washington: National Institute of Justice, February 1987.
- National Institute of Drug Abuse, *National Household Survey on Drug Abuse: Main Findings 1985*, Washington: U.S. Department of Health and Human Services, 1985.

MARKET STRUCTURE AND THE EMERGENCE OF NEW TECHNOLOGIES

R&D Competition for Product Innovation: An Endless Race

By REIKO AOKI*

The purpose of this paper is to explain industry dynamics through the innovative activity of firms. In research-intensive industries, such as pharmaceuticals and high-technology electronics, the constant introduction of new products and the R&D investment to achieve product innovation is critical for the survival of a firm. In such an environment, firms invest in R&D not only to gain immediate profit by selling successful products, but also to maintain the level of their R&D technology or knowledge. A firm's R&D behavior is based on dynamic considerations and this behavior itself will generate the dynamics of the market. In recent years, many traditionally static oligopoly concepts, such as pricing, quantity and investment decisions, have been applied to analysis of industry dynamics in the form of dynamic stochastic games (Lloyd Shapley, 1953). I take a similar approach by extending the patent race models to examine market dynamics. Here the rivalry among firms in an industry continues over many sequences of products and may potentially last forever. There is no predetermined goal or end of the race.

I consider one basic model with two variations. All three are dynamic games with infinitely many discrete periods. In the basic model, a firm's R&D technology is deterministic. Investment always results in progress by one step. Although this is the only model here that is sufficiently tractable for explicit characterization of equilibrium strategies, the system's equilibrium dynam-

ics are limited. In the second model, a stochastic R&D technology is considered. The firm's investment either advances its state of knowledge by one step or leaves it where it was. In the third model, I modify the stochastic R&D so that a firm may advance more than one step. This means that a firm may jump from being a follower to being a leader in one period, without ever being in a tie, that is, a firm may leapfrog (Drew Fudenberg et al., 1983).

In the first model, with deterministic R&D technology, almost all states are absorbing. That is, once those states are reached, the market never changes to another state. The industry evolution literature predicts that asymptotically, the economy converges to concentration or dominance of a single firm. In this model, depending on the initial state of the economy, the economy may asymptotically consist either of multiple firms competing vigorously forever or of just one dominant firm. Also, no matter how low the R&D cost, the industry leader's lead is never large because the follower will drop out when the leader's lead is still small. The introduction of a stochastic element in the second model eliminates the perfect predictability that prevents firms from competing. Although the probability of failure may discourage investment, the positive probability that the leader may fail to make progress causes the follower to remain in the competition for larger leads by the leader. If leapfrogging is possible, the lead may be even greater before the follower drops out. However, even with leapfrogging, the positive cost of investment causes the follower to drop out if the lead becomes too great. Once the follower leaves, the leader stops investing and the game effectively terminates. Thus there is an en-

*Department of Economics, State University of New York, Stony Brook, NY 11794-4384. I am grateful to Beth Allen for her comments that have vastly improved the paper as well as my understanding. I also thank John Hillas for comments.

dogenous end to the game that is reached with probability one and always ends with one dominant firm.

1. The Basic Model

Consider two firms of equal R&D efficiency engaged in R&D competition for an infinite number of periods. During each period, a firm either makes a R&D investment at cost c to advance its state of knowledge by one step, or does not invest anything and remains where it is. Initially, we consider a case where advancement is certain. The firm at the frontier of knowledge sells the current product. New products make previous products completely obsolete so that the firm at the frontier, or the leader during that period, is a monopolist and gets monopoly profit π^m . The follower obtains nothing for the period. If two firms are at the same level of knowledge, products coexist and each firm receives the duopoly profit π^d as its profit for the period. A firm's instantaneous payoff for the period is the profit less investment cost, if any. Investment decisions maximize the discounted sum of instantaneous payoffs given the rival's investment decision. Even with zero profit in the current period, investment may be undertaken in order to improve the state of knowledge which affects future profit prospects.

It is assumed there are no diminishing returns to investment and thus the cost c of investment is independent of the state of knowledge. Profits depend only on the market structure (duopoly or monopoly) and are independent of the size of the lead. These two factors mean that only the *relative* positions matter to the firms and all possible payoff relevant positions of firms (states of the game) can be summarized by the set of integers. The integer k represents the lead of firm 1. If $k < 0$ then firm 2 is the leader, while $k = 0$ means that the two firms are even. Simultaneous investment by firm 1 and no investment by firm 2 increases k by one. If both firms invest, k remains the same since they both advance by one step and their relative position is unchanged. Note also that an infinite sequence of not

investing by the follower is effectively dropping out. The firm then obtains zero profits and expenditure equal zero perpetually, as if firm were no longer competing.

Let us consider an equilibrium in symmetric stationary Markov strategies. A strategy is completely described by specifying the action of each firm at each state k . I use $r^i(k) \in [0, 1]$ to denote the probability of firm i investing at state k . If $r^i(k) = 0$, firm i does not invest, if $r^i(k) = 1$, it does invest, and if $0 < r^i(k) < 1$, firm i plays a mixed strategy at state k . A firm's decision as to what action to take at each state depends only on that state, independent of history up to that point (Markov) and the period in which that state is reached (stationarity). Equilibrium strategies consist of actions that are best responses of each firm given its rival's equilibrium strategy. Thus they are subgame perfect Nash equilibria. For exposition, I focus on states in which firm 1 is at least as advanced as firm 2, that is, on nonnegative values of k (symmetry). An equilibrium associates a pair $(v^1(k), v^2(k))$ to each state k . These are the continuation values of the game and equal the expected discounted sum of instantaneous payoffs when firms invest according to the equilibrium strategies.

The properties of equilibrium differ according to the magnitude of costs relative to profits (see my 1990 paper). First, if cost is high, ($c > (\pi^m - \pi^d)/(1 - \beta)$, where β is the discount factor), neither the leader nor the follower ever invests. The follower does not invest even when it is only one step behind since the stream of profits after getting even is not enough to justify the cost. Because the monopoly profit does not depend on the size of the lead, the leader has no incentive to invest when the follower does not invest. Nor do tied firms invest. The cost is so high that a firm would not choose to take the lead and obtain an infinite sequence of monopoly profits instead of duopoly profits. Thus in an industry where R&D costs are high relative to profit, there is no innovative activity.

On the other hand, if cost is low ($c < \pi^d$), both firms invest when they are even. The value of catching up is less attractive for the

follower and it will drop out even when the lead is small despite the small investment costs. In fact, investment continues only when the follower is one step behind, and at this point the follower is indifferent between investing and not investing. Thus the continuation value for the follower is zero and when he is two or more steps behind, he will not choose to invest—at two or more steps behind, he chooses not to invest as a best response to a leader who is not investing. This is effectively dropping out. When the follower is one step behind, it invests with probability less than one because the low cost means that the leader will invest as a best response to the follower investing. Thus at $k = 1$, firms will play mixed strategies. When the cost is slightly higher (so that it is between the two previous cases), a firm will invest only if the rival does not invest at $k = 0$. Thus, equilibrium differs based on behavior at $k = 0$: neither firm invests, only one firm invests, and both invest, according to the size of investment cost. In all three cases, the follower does not invest at all when it is one or more steps behind ($r^1(k) = r^2(k) = 0, \forall k \geq 1$).

The results of deterministic R&D technology must be interpreted with care. For instance, the case of small cost had an equilibrium of the form $r^1(0) = r^2(0) = 1, 0 < r^1(1) < r^2(1) < 1, r^1(k) = r^2(k) = 0 \forall k \geq 2$. This means that if the economy starts with both firms at the same level, it stays that way forever. States $k \geq 1$ are never reached. If the initial state is $k = 1$, the transition law is stochastic, and with positive probability, the follower may catch up (in which case again both firms invest forever and the two firms stay at that state forever). The leader may also increase the lead to two steps ($k = 2$), in which case the follower ceases to invest and the leader enjoys monopoly profits forever. Similarly, for intermediate cost values, all states except $k = 0$ are absorbing states. Lack of transition from state to state in equilibrium is due to several factors: deterministic technology, zero instantaneous payoff of the follower and the stationarity of strategies.

The remainder of the game is completely predictable except for the stochastic out-

come from mixed strategies. This eliminates a phenomenon such as a follower investing to narrow the lead if the leader ever invests when lead is narrower. If such behavior occurs in equilibrium, the follower knows that it can never get even, which is the first time it obtains positive instantaneous profit. Hence there is no incentive for any investment once it is the follower. Stationarity and zero payoff of the follower eliminates the possibility of returning to a state after several periods with probability one. Since costs must be incurred between the two periods, there is no way the continuation value could be the same, which should be the case for stationary Markov equilibrium. We will never have a situation where firms take turns being leaders and followers.

The deterministic model clarifies the effect of the magnitude of cost. One might expect low costs to accommodate the follower. However, low cost does not imply that the discounted expected sum of profits is high since vigorous competition dissipates profits. The follower still drops out after falling behind by only two steps.

II. A Simple Stochastic R&D Technology

In this section, I introduce uncertainty into the R&D technology. Instead of a firm advancing by one step with probability one when it invests c , stochastic R&D technology means that the advance takes place with probability $p \in (0, 1)$. Thus if both firms invest at state k , the state next period will be $k + 1$ with probability $p(1 - p)$, $k - 1$ with probability $p(1 - p)$ and will remain k with probability $p^2 + (1 - p)^2$. Everything else (instantaneous profits and strategies) remains the same. As with any dynamic stochastic model, the characterization of equilibrium is extremely difficult. One needs to solve both the maximization problem and the continuation value equation for the optimal strategies and continuation values simultaneously for both firms. Note that the action spaces are compact and convex, all instantaneous payoffs are bounded and there are only a countable number of states. Since firms are maximizing the sum of discounted expected payoffs, these condi-

tions guarantee the existence of stationary Markov equilibrium strategies by the standard fixed point arguments (for example, as in A. Federgruen, 1978). I do not attempt a full characterization of equilibrium (strategies and continuation values) here.

We are able to answer the following question: will the follower that is two or more steps behind try to catch up when the technology is stochastic? Stochastic R&D technologies eliminate the complete predictability which prevented the follower from trying to eliminate the lag in the deterministic model. Of course the follower's chance of successfully doing so is adversely affected by the R&D technology. On the other hand, with deterministic R&D, when the cost was low, the leader's best response to follower investing was to invest. However, with stochastic R&D, for the same range of cost, a leader may choose not to invest if the probability p of success is small. This is because the chance that the follower will not succeed in advancing reduces the leader's incentive, thereby favoring the follower. The net effect is that the follower becomes more aggressive for very low values of p , and will not drop out even after it is two steps behind, if cost is sufficiently small.

III. The Possibility of Leapfrogging

I now consider a stochastic R&D technology with leapfrogging (see my 1988 paper). The cost of investment is still constant c , but now a firm may advance by zero or any positive number of steps. Smaller advances are more likely than larger advances. As before, no investment means no advancement with probability one. All instantaneous profits and strategies remain the same.

As in the deterministic case, if c is small both firms will invest when the two firms are tied. There actually may be a range of small leads (small k 's) where both leader and the follower invest ($r^1(k) = r^2(k) = 1$). Again, since the monopoly profit does not depend on how many steps ahead the leader is, leader will never invest if the follower is not investing. Despite the possibility of leapfrogging, there are still leads large

enough for the follower to stop investing. Together with the preceding observation, this implies that there will be a maximum difference between leader and follower. With leads greater than the maximum, the follower will drop out and the game effectively ends. Thus if a firm is lucky and attains a very large lead, it will remain a monopolist forever without further effort.

The possibility of leapfrogging means that a follower this period has a positive probability of gaining a large enough lead next period to force the present leader out. This may cause the follower to continue investing even when the lead is two or more steps. On the other hand, the positive probability that the follower may not catch up may cause the leader to choose not to invest. Thus there may be a nonempty set of states where the follower invests with probability one ($r^2(k) = 1$) and the leader invests with probability zero ($r^1(k) = 0$). This also implies that the continuation value of the follower may not be monotonic, contrary to the deterministic model above and other patent race models (see Christopher Harris and John Vickers, 1987; Steve Lippman and Kevin McCardle, 1987). Consider a state \hat{k} where $r^2(\hat{k}) = 1$ and $r^1(\hat{k}) = 0$. The follower invests but the outcome of R&D is such that it remains a follower but now the lead k' is smaller, so that $r^1(k') = r^2(k') = 1$. Fixing the action of the leader, the follower would prefer to be at k' . However, the follower prefers the action of the leader at \hat{k} . It may be that this second effect dominates. The follower may find itself in a worse position at k' than at the larger \hat{k} with the continuation value smaller at k' than at \hat{k} . In this case there is a nonmonotonicity in the continuation value of firm 2. When k is very negative, firm 2 is a monopolist and the continuation value is at its maximum. As k increases, the continuation value decreases. However, the continuation value increases between k' and \hat{k} .

IV. Concluding Remarks

I have examined three different R&D technologies: deterministic, stochastic without leapfrogging, and stochastic with leap-

frogging. Although the deterministic model illuminates the relevant tradeoffs as well as the effects of R&D costs, in equilibrium there are almost no real dynamics. Initial states are almost always the asymptotic states as well. The lack of dynamics is partially due to the restriction to stationary Markov strategies. On the other hand, my results suggest that one explanation for a single firm's dominance in an industry is the deterministic nature of R&D. Success is possible with probability one to the follower as well as the leader, but the follower never tries to catch up because it knows the leader will then invest.

Even with single-step advances, stochastic R&D prolongs competition. That is, the follower may still continue to invest even when it is two steps behind. Uncertainty of success works to the follower's advantage by reducing the leader's incentive to invest when the follower is investing. The shortcoming of this approach is the difficulty of completely characterizing equilibrium. Leapfrogging facilitates the characterization because the set of possible states in the next period is larger from any state. In particular, there is always positive probability of entering states with positive instantaneous payoffs if a firm invests, even when it is the follower. In equilibrium, there will be actual transitions between various states, while monopoly with the follower inactive will be the only absorbing state.

There are a few simplifications I have made which are crucial to the results. Instantaneous profit is independent of how large the lead is. The state of knowledge of a firm would typically be reflected in the quality of product the firm sells. There are cases where the profit would depend not only on the *relative* quality but also on *absolute* quality. If that were the case, the leader would have an incentive to continue improving even after it has become a monopolist, but the result that the follower drops out

when the lead is too large still should be true. Moreover, with heterogeneous products or consumers, the follower's instantaneous profit may be positive. This would give the follower an incentive to continue investing even when it is far behind, so that it may not ever drop out.

Although there is exit by a firm in my model, there is no entry. The lack of possibility of entry allows the leader to be a monopolist forever without any investment in technology. In the stochastic models, monopoly with the follower inactive is the only absorbing state and asymptotically the market becomes a monopoly with probability one. Yet, to describe industry evolution, entry should be considered. Otherwise, one firm dominance occurs almost by definition.

REFERENCES

- Aoki, Reiko, "R&D Rivalry Over Time: A Dynamic Stochastic Game Approach," Ohio State University, mimeo., 1988.
- , "Product Innovation with Deterministic R&D Technology," SUNY-Stony Brook, mimeo., 1990.
- Federgruen, A., "On N-person Stochastic Games with Denumerable State Space," *Advances in Applied Probability*, No. 2, 1978, 10, 452–71.
- Fudenberg, Drew et al., "Preemption, Leapfrogging, and Competition in Patent Races," *European Economic Review*, June 1983, 22, 3–31.
- Harris, Christopher and Vickers, John, "Racing with Uncertainty," *Review of Economic Studies*, January 1987, 54, 1–21.
- Lippman, Steve A. and McCardle, Kevin F., "Dropout Behavior in R&D Races with Learning," *RAND Journal of Economics*, Summer 1987, 18, 287–95.
- Shapley, Lloyd, "Stochastic Games," in *Proceedings of the National Academy of Science*, 1953, 39, 1095–100.

Choosing R&D Projects: An Informational Approach

By BETH ALLEN*

This paper argues that research and development (R&D) should be viewed explicitly as information acquisition activities. A firm's choice among various R&D strategies is equivalent to deciding what information to gather. Moreover, the pursuit of different research projects involves searching for different sorts of information.

Thus, R&D is not homogeneous. To attempt to summarize R&D activities by a single aggregate measure (for example, by the cost of R&D inputs or by modeling the intensity of only a single exogenously given project) is inherently misleading. Such a story misses the important possibilities for alternative R&D projects to serve as complements or substitutes for each other. The neglect of heterogeneity risks misleading conclusions about welfare effects based on underinvestment or overinvestment in research and development.

This paper proposes an alternative abstract analysis of R&D activities that focuses on filling the gaps discussed above. I suggest that the successful outcomes of projects can be specified by differentiated information. Each project thus is represented by a probability of success (or an arrival rate for the dynamic version of our model) as well as a description of the knowledge that may be acquired. This formulation facilitates the comparison of projects in terms of substitution, complementarities, and dominance relationships (where one project dominates another if it yields superior infor-

mation with at least the same success rate or equivalent information with a strictly larger probability or arrival rate).

Despite the preceding observations, the literature emphasizes homogeneous or aggregate R&D activity. An exception is provided by the portfolio choice analogy of Sudipto Bhattacharya and Dilip Mookherjee (1986). They permit different risk profiles when different research methods are employed to search for the same discovery. The informational content of all successful projects is the same; only their riskiness differs. The authors analyze the effects of mean-preserving spreads and correlation on the choice of privately and socially optimal R&D investment decisions with risk aversion. Kenneth Judd (1986) analyzes a multi-stage game in which firms choose R&D portfolios. Partha Dasgupta and Eric Maskin (1987) consider a model in which each firm chooses to locate its research project in an interval of parameters. Dasgupta (1990) examines the choice of an endogenous (finite) number of independent and identical "parallel" projects. Finally, Richard Nelson (1982) discusses complementarities among projects and externalities from research to development; an endogenous (finite) number of parallel research projects are undertaken followed by development of the best project.

I. A Model of Heterogeneous R&D Projects

This section summarizes a microeconomic model of stochastic information acquisition as a result of alternative R&D activities. At this point, before any notation is introduced, clarification is needed for the fundamental concepts of information, R&D projects, and success. Information (sometimes termed an information structure for emphasis) is a generalized partition of states of the world. Here information is an *ex ante* concept (i.e., the ability to know that the true

*Milton C. Denbo Term Professor of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297. This research was supported by NSF grant No. SES88-21442. I thank the participants, especially Reiko Aoki, in the Industrial Organization Program (organized by Suzanne Scotchmer and Yair Tauman) of the 1990 Stony Brook Summer Institute on Game Theory and Economics for helping to educate me about the recent R&D literature. Pamela Brown and Tara Vishwanath provided helpful comments on the preliminary version of this paper.

state of the world lies in one of the subsets defining the partition) rather than the *ex post* fact that the true state actually lies in a certain subset. Success means that such an information structure is acquired and can be used to condition future economic decisions. An R&D project is defined to be an attempt to obtain access to a particular information structure. Specification of a project and its intensity yields a success rate, either in the form of the probability of success or an arrival rate for success.

Perhaps an analogy with research in economics would be useful. Let us begin with a specific question (think of this as the information we wish to obtain), such as whether a class of models has an equilibrium. Success is interpreted as the completion of this piece of economic research. For instance, we find a counterexample to the existence of equilibrium, we prove that equilibria exist, or we find additional (necessary) conditions (for example, strictly concave and twice continuously differentiable utilities) under which we can demonstrate that there is an equilibrium. Notice that success could involve any of these three definite outcomes or sets in the partition. In particular, discovery of a counterexample constitutes success, while lack of success means that we have been unable to reach any clear conclusion about the equilibrium problem for our class of models. Finally, in this context, a research project is determined by the selection of an open question and an "intensity" of research (which would, in fact, correspond to the difficulty of the various methods of attack that we examine). When a project succeeds, we clearly stop working on that particular project, although we may or may not (i.e., if the result is an uninteresting counterexample) publish the results.

To return to the formal model, let Ω be the set of states of the world, and think of Ω as sufficiently large to incorporate all of the economically relevant uncertainty except the random success or failure of research projects.¹ Typical states of the world will be denoted $\omega \in \Omega$. A measurable event

is a subset of Ω that economic agents "eventually learn." Let \mathcal{F} be the σ -field of measurable events. Then define a probability measure μ on the measurable space (Ω, \mathcal{F}) so that the abstract probability space $(\Omega, \mathcal{F}, \mu)$ summarizes all of the uncertainty in the world except project successes. An interpretation is that knowledge of ω (or of an event in \mathcal{F}) tells us all of the technologies that could ever be discovered.

As indicated earlier, information is taken to be a generalized partition (defined by subsets in \mathcal{F}) of Ω . Formally, an information structure is a complete sub- σ -field of \mathcal{F} .² Let \mathcal{F}^* denote the set of complete sub- σ -fields of \mathcal{F} and make \mathcal{F}^* into a metric space via the distance function

$$d(G, H) = \sup_{G \in \mathcal{G}} \inf_{H \in \mathcal{H}} \mu(G \Delta H) \\ + \sup_{H \in \mathcal{H}} \inf_{G \in \mathcal{G}} \mu(G \Delta H)$$

where \mathcal{G} and \mathcal{H} belong to \mathcal{F}^* and, for $G \in \mathcal{F}$ and $H \in \mathcal{F}$, $G \Delta H = (G \cup H) \setminus (G \cap H)$ is their symmetric difference. (See my 1990 paper and the references listed there for details.) The set \mathcal{F}^* of information structures represents all possible research questions.

To define R&D projects, we must also specify their intensities (interpreted as summary measures of the research methodology in terms of its corresponding costs). This is done by associating a nonnegative number with an information structure. Hence the ordered pair (G, c) , where $G \in \mathcal{F}^*$ and $c \in \mathbb{R}_+$, denotes a research project. Generalization of the model to dynamic R&D projects in which the intensities vary over time is obvious.

Uncertain R&D projects require also that success rates be added to the model. These success rates are completely specified by the research project; that is, by the choice of an information structure and its intensity. To

¹Later I discuss the removal of this qualification.

²This is needed so that if the distance between two information structures is zero, then they are the same. A complete σ -field is one that contains all subsets of sets of measure zero. Of course, we also need completeness of \mathcal{F} .

simplify some mathematical technicalities, I shall assume independence (for now) and ignore the measurability problem. Thus, write $\pi: \mathbf{F}^* \times \mathbb{R}_+ \rightarrow [0, 1]$ for the success rate of an R&D project in the static version of the model, where $\pi(\mathbf{G}, c)$ gives the probability that project (\mathbf{G}, c) succeeds. In the dynamic model, let $\lambda: \mathbf{F}^* \times \mathbb{R}_+ \rightarrow [0, 1]$ give the (Poisson) arrival rate for success, where $\lambda(\mathbf{G}, c)$ defines the success arrival rate of project (\mathbf{G}, c) . I assume that all economic agents know the success rates or success arrival rates of all R&D projects.³

The value of a project is the amount by which its expected rewards W exceed its cost c . For the static model, write the value function V as

$$V(\mathbf{G}, c; \pi(\mathbf{G}, c)) = \pi(\mathbf{G}, c)W(\mathbf{G}) - c$$

while, in the discrete time dynamic model, the value V_t at time t is

$$\begin{aligned} V_t(\mathbf{G}, c; \lambda(\mathbf{G}, c)) \\ = w_t(\mathbf{G})\Pr\{\text{success at or before } t | \lambda(\mathbf{G}, c)\} \\ - c \Pr\{\text{no success before } t | \lambda(\mathbf{G}, c)\} \end{aligned}$$

where $w_t(\mathbf{G})$ denotes the project's expected return (if successful) in period t ,⁴ so that the expected discounted net present value (with discount rate δ) is

$$\begin{aligned} V_\delta(\mathbf{G}, c; \lambda(\mathbf{G}, c)) \\ = \sum_{t=1}^{\infty} \delta^t w_t(\mathbf{G}) \Pr\{\text{success at or before } t | \lambda(\mathbf{G}, c)\} \\ - \sum_{t=1}^{\infty} \delta^t c \Pr\{\text{no success before } t | \lambda(\mathbf{G}, c)\} \end{aligned}$$

³If this is not satisfied, then the two-armed bandit problem may arise and firms could choose strictly dominated projects infinitely often. See, for instance, Michael Rothschild (1974).

⁴ $w_t(\mathbf{G})$ could depend on t if, for example, population growth affects the size of the market for the innovation.

or, alternatively, if t^* denotes the random time at which success occurs,

$$\begin{aligned} V_\delta(\mathbf{G}, c; \lambda(\mathbf{G}, c)) \\ = E_{t^*} \left[\sum_{t=t^*}^{\infty} \delta^t w_t(\mathbf{G}) - \sum_{t=1}^{t^*} \delta^t c \right]. \end{aligned}$$

Observe that it is assumed that costs are paid during every period up to and including the time when success occurs (or the discovery "arrives"). The formulae for the dynamic case could easily be modified to permit a sequence of discoveries, so that a finite or infinite sequence of increasing information σ -fields arrive at dates given by a sequence of random times.

Since additional information can always be ignored by the firm, the reward functions are nondecreasing; that is, if $\mathbf{G} \supseteq \mathbf{H}$, then $W(\mathbf{G}) \geq W(\mathbf{H})$ and $w_t(\mathbf{G}) \geq w_t(\mathbf{H})$. This implies that we have $V(\mathbf{G}, c; \pi(\mathbf{G}, c)) \geq V(\mathbf{H}, c; \pi(\mathbf{H}, c))$ whenever $\mathbf{G} \supseteq \mathbf{H}$ and $\pi(\mathbf{G}, c) \geq \pi(\mathbf{H}, c)$. Moreover, for the dynamic case, $\mathbf{G} \supseteq \mathbf{H}$ implies $V_t(\mathbf{G}, c; \lambda) \geq V_t(\mathbf{H}, c; \lambda)$ and $V_\delta(\mathbf{G}, c; \lambda) \geq V_\delta(\mathbf{H}, c; \lambda)$.

I assume that the value functions are jointly continuous in their arguments. This is innocuous when information is used to maximize a conditional expectation and/or when information structures determine the firm's production function in a continuous way. (See my 1983 article for a formal statement and proof.) On the other hand, concavity cannot hold because our underlying set of information structures does not have the structure of a convex set.

Finally, note that the success rate π and success arrival rate λ specifications above implicitly assume that the outcome of each project undertaken by a firm does not depend on the other R&D projects in the profile chosen by the firm. Moreover, success was assumed to be independent of the underlying basic state of the world, $\omega \in \Omega$. At the expense of more complicated notation and additional mathematical difficulties, we could incorporate R&D project successes into the description of the set Ω of states of the world. Then a generalized random variable defined on Ω would indicate which projects would succeed if they were

to be performed. Similarly, we could define our success random variable so that it depends on the entire set of R&D projects that are actually undertaken.

II. The Firm's Choice Problem

Assume either that the firm is risk neutral (so that it maximizes expected profits) or risk averse with a utility function (defined over profits) that is continuous, increasing, and concave. Then the firm's optimal (static or dynamic) choice of a single R&D project among a compact set of projects is well defined, but need not be unique; maximizers exist but ties may occur.

If the firm's possibilities are expanded to encompass the selection of several projects that may serve as complements or substitutes for each other (think of competing research teams within the firm), a generalization of the firm's choice set from projects to (finite and bounded) sets of projects is needed.⁵ To do this, represent each project by the probability measure that assigns mass one to it. A profile or portfolio of projects is given by a finite sum of such probabilities. This method yields a compact choice set so that the maximum theorem applies (because the firm's objective function depends continuously on research profiles). (Technical details are in my 1986 paper.)

III. Strategic Interaction

Section II considered the behavior of a single firm in its choice among R&D activities. Here I wish to permit strategic interaction among competing firms. The natural way to do this is to analyze the noncooperative game among firms in which strategies are sets of projects.

Unfortunately, whenever firms choose among R&D projects (even if they are restricted to finitely many alternative projects), the game does not generally have a Nash equilibrium in pure strategies. With a finite

number of pure strategies, convexity of (pure) strategy sets fails, but the problem may be solved by resorting to mixed strategies. For the choice problem involving (bounded) profiles of R&D projects (from an arbitrary compact set of information structures) described earlier, again the extension to mixed strategies restores convexity of strategy sets and best-reply correspondences. The usual problems arise here in the interpretation of mixed strategies. Perhaps the best story in this context is to imagine each firm optimizing against its probabilistic beliefs about other firms' behaviors.

Continuous dependence of payoffs on the strategy choices of other firms can be derived from the stochastic nature of R&D success. In other words, appropriate choices of the π and λ functions yield expected payoffs which exhibit continuity in the (pure or mixed) strategy profiles of all firms.

IV. Welfare Consequences

For the important issue of whether underinvestment or overinvestment in research and development occurs, the hypothesis of homogeneous R&D is especially troublesome. The classical literature (i.e., Kenneth Arrow, 1962, and Jack Hirshleifer, 1971), as well as more contemporary treatments, analyzes the problem by comparing the actual level of R&D investment with the socially optimal amount of aggregate information acquisition activities.

If every project yields exactly the same innovation when it succeeds, patent races tend to create excessive waste. On the other hand, duplication of effort is unlikely to be socially suboptimal if firms pursue distinct projects or if intermediate discoveries can be shared. Yet these aspects can be explicitly modeled only when differentiated information is incorporated.

For spillovers, a similar problem arises. A spillover frequently takes the form of information obtained from a research project oriented to one purpose that can then be joined together with additional information from another research project undertaken by a different firm. Consequently, the com-

⁵For the multi-armed bandit problem with parallel projects, Tara Vishwanath (1990) shows that the number and type of projects chosen may change over time.

bination is utilized for a different purpose. Such an analysis demands that different types of R&D projects be explicitly considered.

Finally, the recent public policy discussions of cooperative R&D ventures provide another example where heterogeneity of information is relevant for welfare conclusions. Surely the most interesting cases of potentially beneficial cooperation are based on economies of scope and specialized expertise of participants. These factors may be fully included in the analysis only when appropriately described differentiated R&D projects constitute allowable strategies.

REFERENCES

- Allen, Beth, "Neighboring Information and Distributions of Agents' Characteristics Under Uncertainty," *Journal of Mathematical Economics*, September 1983, 12, 61-101.
- , "The Demand for (Differentiated) Information," *Review of Economic Studies*, July 1986, 53, 311-23.
- , "Information as an Economic Commodity," *American Economic Review Proceedings*, May 1990, 80, 268-73.
- Arrow, Kenneth J., "Economic Welfare and the Allocation of Resources for Invention," in R. Nelson, ed., *The Rate and Direction of Inventive Activity: Economic and Social Factors*, Universities-NBER Conference Series, No. 13, University Microfilms, 1962.
- Bhattacharya, Sudipto and Mookherjee, Dilip, "Portfolio Choice in Research and Development," *RAND Journal of Economics*, Winter 1986, 17, 594-605.
- Dasgupta, Partha, "The Economics of Parallel Research," in Frank Hahn, ed., *The Economics of Missing Markets, Information, and Games*, Oxford: Clarendon Press, 1990, ch. 6.
- and Maskin, Eric, "The Simple Economics of Research Portfolios," *Economic Journal*, September 1987, 97, 581-95.
- Hirshleifer, J., "The Private and Social Value of Information and the Reward to Inventive Activity," *American Economic Review*, September 1971, 61, 561-74.
- Judd, Kenneth L., "Closed-Loop Equilibrium in a Multi-Stage Innovation Race," mimeo., Hoover Institution, Stanford University, 1986.
- Nelson, Richard R., "The Role of Knowledge in R&D Efficiency," *Quarterly Journal of Economics*, August 1982, 97, 453-70.
- Rothschild, Michael, "A Two Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, September 1974, 9, 185-202.
- Vishwanath, Tara, "Parallel Search for the Best Alternative," mimeo., Department of Economics, Northwestern University, 1990.

The Determinants of Investment in New Technology

By SARAH J. LANE*

New technologies displace older methods and lead to improvements in productivity. Since the decision to invest in a new technology depends on the costs and benefits of adoption, the use of new technologies will not be uniform across either technologies, adopters, or time. This paper extends the previous literature on the diffusion of new technology by examining the effect of the structure of the output market and the effect of investment in earlier competing technologies on the decision to adopt. The case studied in detail is the use of continuous mining machinery in underground bituminous coal mining.

Neoclassical models assume that the integrated firm has the authority to make employment, production, distribution, and investment decisions for all divisions of the firm. As a result, vertical integration may alter control over decision making, and a vertically integrated firm may have different rates of technological adoption. Indeed, I find that vertical integration is an important determinant of the rate of adoption. I also conclude that the installed base of earlier competing technologies affects the adoption of new technology.

Bituminous coal mining in the United States is well suited for such a study since the market structure of the coal mining industry has large variation; along with thousands of very small producers, there are a number of large producers. In addition, there is also variation in the structure of the output market, since a subset of mines are integrated with their consumers such as steel companies or electric utility companies.

Like E. Mansfield (1961), this paper looks at the costs and benefits of the introduction

of a new technology. It differs from previous work by estimating a stock-adjustment model so as to take advantage of the state-level panel data and to allow for shifts in the rate of adoption over time.

I. Coal Mining in the United States: 1945-75

The trends in production have been a movement towards the West from the East and towards surface from underground mining. Only those states with a significant portion of underground production are included in the study (Alabama, Colorado, Illinois, Indiana, Kentucky, Maryland, Ohio, Pennsylvania, Tennessee, Utah, Virginia, and West Virginia.)

Underground mining operations consist of cutting, drilling, blasting, loading, and hauling the coal. Traditionally, these procedures were carried out in separate, sequential operations. Innovation began with the mechanization and improvement of each step in sequence to eliminate successive bottlenecks. More recent innovations, such as continuous mining machines and longwall mining machines, integrated the separate steps of the coal mining process. In continuous mining, a rotating head digs into the coal seam, removes the coal, and then transfers it to a conveyor for transport out of the mine. Coal was mined using this method beginning after World War II, and accounted for over 50 percent of coal output by the 1970's. Longwall mining machines added roof supports but have not been widely introduced. Since continuous mining machines replace a sequence of machines, the installed base of competing technology would be expected to affect the adoption of the new innovation.¹

*Assistant Professor, School of Management, Boston University, Boston, MA 02215. I thank Greg Clark, Zvi Griliches, Joyce Jacobsen, Leslie Papke, Robin Prager, Joshua Rosenbloom, and John Strudwick. John Pencavel provided access to part of the data. Boston University provided funding for the collection of additional data.

¹For example, G. E. Hale (1970), suggests that Consolidation Mine was slow in adopting the continuous miner partly because it had already modernized its digging and loading facilities with less advanced devices.

Sales of coal can take place in the open market or in the captive market. Captive production is that portion of production that is sold to or transferred to the parent company or a subsidiary of the parent company. The share of captive coal has been very stable over time at 16–17 percent on average. Illinois, Indiana, Maryland, Tennessee, and West Virginia have very low levels of captive production (approximately 2 percent), while Alabama, Pennsylvania, and Utah have high levels (approximately 40 percent).

If vertical integration does indeed alter decision making, the expected direction of the effect would be to increase the rate of adoption. As suggested by the literature, the vertically integrated firm may have a more stable demand. The development of a mine requires investment in mine-specific assets. Optimal technology is chosen to meet expected demand, and if that demand is more certain, as in the case of vertical integration, then innovation may be more likely.²

II. The Empirical Framework

Diffusion of an adoption is based on a decision to invest in new technology. Given this connection between investment and diffusion, the approach used is to estimate a basic stock-adjustment model where the proportion of production using the new technology approaches the desired proportion at some rate β .

$$(1) \quad \delta Pr_{it} / \delta t = \beta_{it} (Pd_{it} - Pr_{it-1}) \\ = \beta_{it} Pd_{it} - \beta_{it} Pr_{it-1},$$

where Pr_{it} is the proportion of output in state i at time t using the new technology, β_{it} is the rate of adjustment, and Pd_{it} is the desired proportion or "ceiling" level. Both the rate of adjustment and the desired proportion vary over time and across states.

²For example, consider this statement from ARCO: "In our underground mines... we have the opportunity to invest in an efficient longwall unit.... Our investment is contingent upon us securing long-term contracts, which we are currently negotiating" (p. 11, ARCO Security Analyst Meeting, March 20, 1990).

The variables that are expected to determine the ceiling reflect primarily technical constraints, and the variables that are expected to determine the rate reflect primarily economic constraints.

This stock-adjustment approach uses the full information available in the panel data set. Previous studies have used a logistic model that, because it places primary emphasis on the role of time in diffusion, assumes that potential adopters have an equal and time-invariant probability of adoption. The stock-adjustment model, in contrast, allows for shifts in the rate of adoption over time in response to changes in the economic costs and benefits of adoption.

I use the percentage of production using the new technology as the measure of adoption, since it is an accurate and unbiased measure of the production experience arising from the adoption decision. Since both mine size will vary and there will be a mix of technologies in any one mine, a measure based on the number of mines may severely over- or underestimate the coal mined using the new technology. For instance, by 1960, 47 percent of underground mines were using continuous mining machines, but the output mined using these machines accounted for only 27.7 percent of total underground production.

While the mine would be the ideal unit of analysis, only state-level data are available. Given this level of aggregation, entry and exit of mines will affect both the desired proportion of production using the new technology and the measured rate of adoption. An apparent increase in production will occur if the nonadopters are the older or marginal mines that eventually exit or fail. Similarly, an increase will occur if entering mines use the new technology and if adopters grow faster than nonadopters. This possible bias will exist with *any* adoption or diffusion measure when firms are entering or exiting the sample over the period. However, there is clearly economic content to the information that mines or firms that have not adopted are leaving the sample, and that firms entering the sample or growing faster are using the new technology. Therefore, the net change in the number of mines is included as an explanatory variable

TABLE 1—EXPLANATORY VARIABLES^a

	Ceiling	Rate	
A. Data that vary with both state and time are			
<i>AVGM</i>	(+)	(+)	Average underground mine size (in millions of tons)
<i>CMPM</i>	(-)		Number of cutting machines per mine
<i>MLN</i>	(-)		Proportion of underground production mined using mechanical loading machines ^b
<i>NUMMP</i>		(+)	Percentage change in the number of underground mines
<i>NSOMP</i>		(+)	Percentage of production not sold on the open market
<i>RESM</i>	(+)		Reserves per mine (total reserves of state in millions of tons/number of mines) ^c
<i>SW</i>	(+)		Average seam width ^c
B. Data that vary over time but not by state are			
<i>STOCK</i>		(-)	Total stock of coal in the economy (in millions of tons)
<i>WAGES</i>		(+)	Hourly wage in bituminous coal mining deflated by the CPI
<i>TIME</i>		(+)	Time trend

^aExpected signs are shown in parentheses.

^bSince the data on mechanical loading include production loaded by continuous mining machines, this variable is calculated as the percentage of production mechanically loaded minus the percentage of production mined by continuous mining machines.

^cFor these variables, data are not available for all years. The closest preceding year's data were used when observations were missing.

(gross data are not available). All the explanatory variables are defined in Table 1.

A measure of mine size is included since there is a greater probability that large mines will need to replace older equipment at any point in time. In addition, large mines are likely to have a wider range of operating conditions, making it more likely that they have a set of economic and geological conditions suitable for the new technology. Finally, there may be some minimum efficient mine size for the new technology.

Coal reserves are the stocks in situ, and are defined in terms of the short-term economic feasibility of extraction. Since there will be some minimum amount of reserves

necessary to justify the cost of the machines, reserves will negatively affect the ceiling. The width of seams is a measure of the feasibility and ease of introducing continuous mining. Continuous mining machines use less, though possibly more highly skilled labor. The net effect of an increase in wages should increase the rate of adoption if the labor-saving effect outweighs the skill effect. An increase in stocks which decreases the expected near term future price has a negative expected effect.

Finally, I include a time trend to capture all the time-varying variables that do not differ by state. Time also captures the effect of decreased uncertainty about the reliability and costs of using the new technology from observing earlier adopters. It should therefore have a positive coefficient if this effect is dominant.

III. Results and Conclusions

While the logistic model does not reveal much information about the underlying forces affecting the adoption decision, it does summarize the time-series of adoption. A logistic model estimated by state fits the data reasonably well for all states except Alabama and Indiana.³ In these states, after an initial period of an increase, production using this method drops dramatically as mines exit. Using the logistic estimates to calculate the year at which 50 percent of production is mined using continuous mining machines, I find that the three earliest adopters are also states that have relatively high levels of captive production. A similar though less pronounced pattern exists for the slower adopters. These results suggest that the state variation in captive sales is potentially an important determinant of the adoption of continuous mining machines.

Table 2 presents results for the stock-adjustment model estimated using the panel data. The estimation is a nonlinear function of the explanatory variables since the rate interacts with the ceiling proportion. The

³A more detailed version of this paper, available from the author, contains the estimates for the logistic.

TABLE 2—ESTIMATED ADJUSTMENT MODEL—
CONTINUOUS MINING MACHINES ($N = 228$)^a

Coefficients	(1)	(2)	(3)
A. Rate			
Constant	-0.044 (-0.347)	-1.325 (-5.817)	0.054 (1.091)
<i>NSOMP</i>	0.009 [0.106; 0.186]	0.736 (5.960)	0.030 (1.061)
<i>AVGM</i>	0.111 [0.159; 0.226]	0.762 (5.338)	-0.114 (-1.926)
<i>NUMMP</i>	-0.041 [-0.062; 0.175]	-0.050 (-0.812)	0.038 (1.280)
<i>STOCK</i>	0.396 [0.082; 0.012]	3.816 (1.641)	0.333 (0.879)
<i>WAGES</i>	0.008 [3.907; 0.450]	0.261 (2.882)	-0.034 (-1.183)
<i>TIME</i>	0.001 [20.368; 5.900]	0.010 (2.047)	0.001 (0.898)
<i>RESM</i>	-0.007 [1.010; 1.942]	0.013 (2.840)	0.032 (1.802)
B. Ceiling			
Constant		98.582 (56.261)	-289.629 (-1.330)
<i>AVGM</i>		-3.224 (-2.692)	695.182 (1.784)
<i>SW</i>		0.405 (2.041)	40.304 (1.692)
<i>MLN</i>		-1.090 (-46.702)	
<i>CMPM</i>		2.819 (3.958)	-17.564 (-1.429)
Log Likelihood	-783.740	-588.815	-704.782
R-squared	0.029	0.824	0.547

^aDependent Variable[2.838; 7.651]. *T*-statistics are shown in parentheses. The mean and standard deviations of the variables are shown in square brackets.

full set of data is available for the years 1955–66 and 1968–74.

Only average mine size is (marginally) significant in the specification that sets the ceiling proportion constant at 100 (col. 1). Allowing the ceiling to vary (cols. 2 and 3) improves the estimation. Whenever *MLN* is excluded (as in col. 3), the signs and the significance of rate coefficients change dramatically and are no longer consistent with a priori expectations. With this one exception, including or excluding additional explanatory variables for the rate or ceiling has little effect on the estimation results.

The estimated ceiling proportion at the mean of the explanatory variables is 45 percent for column 2. As expected, the proportion of output using the competing technol-

ogy, *MLN*, has a negative effect on the ceiling, and seam width has a positive effect. The ceiling proportion also varies positively with the number of cutting machines. Cutting machines are an older technology, and therefore this variable may be capturing the age of the equipment in the mine instead of the effect of competing technology.

The estimated value for the rate when all the explanatory variables are at their mean levels is .44 for column 2. The signs of the significant variables in column 2 do accord with a priori expectations and also indicate that the structure of the output market does indeed affect the decision to adopt new technology. Captive mines switch to production using continuous mining machines at a faster rate. Higher wages and a higher level of reserves lead to faster adoption, and the rate also increases with time.

Estimating an adjustment model so as to take advantage of panel data and to allow for shifts in the rate of diffusion over time can identify factors that affect the rate of adoption. I conclude that the installed base of earlier competing technologies affects the ceiling or desired proportion of production using the new technology, and through this the adoption of the new technology. The results also highlight the importance of the relationship between investment decisions and vertical integration. In a long-term stable contractual arrangement, such as vertical integration, firms are more likely to adopt firm-specific new technologies.

REFERENCES

- Hale, G. E., "The Case of Coal: Should All Horizontal Mergers Be Held Illegal?," *Journal of Law and Economics*, October 1970, 13, 421–37.
- Mansfield, E., "Technical Change and the Rate of Imitation," *Econometrica*, October 1961, 29, 741–66.
- National Coal Association, *Bituminous Coal Data*, Washington, D.C., annual.
- _____, *Bituminous Coal Facts*, Washington, D.C., annual.
- U.S. Department of the Interior, Bureau of Mines, *Minerals Yearbook*, Washington: USGPO, annual reports for 1945–76.

Diversification by Regulated Monopolies and Incentives for Cost-Reducing R&D

By KAREN PALMER*

Traditional cost-based regulatory methods fail to promote static efficiency in production and do not provide incentives for firms to invest in process innovations that lead to lower production cost in the future. (See Sanford Berg and John Tschirhart, 1988, ch. 10.) Production efficiency may be enhanced if a regulated firm is allowed to exploit economies of scope by diversifying into other unregulated lines of business. If the regulator is able to reduce the regulated revenue requirement due to the cost savings resulting from economies of scope, then some of the benefits of joint production ultimately may be passed on to the consumers of the regulated products in the form of lower prices. Diversification into unregulated markets also increases the firm's incentives to engage in process innovation by providing opportunities for earning profits in unregulated markets. Investments in process innovation can lead to lower future costs of producing the regulated product and potential future price reductions to regulated customers.

I present a model of the firm's decision to diversify into an unregulated market and look at the implications of that decision for a firm's incentive to innovate and for consumer welfare. The firm's choice is strongly influenced by the extent to which the regulator reduces the amount of fixed cost included in the regulated revenue requirement should the firm choose to diversify. I show that if the regulator is able to observe production costs accurately at the time of a rate review, then diversification increases the firm's incentive to innovate and leads

to higher gains in consumers' welfare over time.

I. The Model

I use a three-period model to compare the profit-maximizing behavior of an undiversified regulated firm and a diversified regulated firm. It is assumed that the demand functions for both the regulated and unregulated products do not change over time. The technology used by the firm exhibits economies of scale in the production of the regulated product and economies of scope in the joint production of the two products. The firm is regulated under zero-profit regulation with a one-period regulatory lag; regulators do not adjust prices in response to changes in production costs until the subsequent period. The firm decides whether or not to diversify and how much to invest in R&D in period 1. The full cost reductions from the R&D investment occur in period 2, continue through period 3 and are known with certainty to the firm, but not to the regulator, at the time of the R&D investment. The regulated price is set equal to average production cost (not including R&D expenditures) in the regulated market in period 1 and remains the same through period 2. Price is adjusted in period 3 to yield zero profit on sales of the regulated product. Also assume that the regulator can observe at each rate review all the cost information necessary to set a price that yields zero profit and that the regulatory constraint is binding during the period when a rate review occurs.

If the regulated firm decides not to diversify in period 1, then the present discounted value of its profits over the next 3 periods is

$$(1) \quad \underset{B}{\text{Max}} -B + \delta [\text{Rev}(x_2^u) - \gamma(B) \times [F + C(x_2^u)]] + 0$$

*Fellow, Resources for the Future, 1616 P St. NW, Washington, D.C. 20036. I thank Mike Toman, Dallas Burtraw, Molly Macauley, Steve Polasky, and Tim Brennan for helpful comments on an earlier version of this paper.

where δ is the discount factor, $\text{Rev}(x_2^u)$ is the revenue from selling x in period 2, B is the level of investment in process innovation, and $\gamma(B)[F + C(x)]$ is the cost function for x .

The cost function for x has a fixed component, F , and a variable component $C(x)$ with $C_x > 0$ and $C_{xx} < 0$. Investment in R&D in period 1 reduces both components of cost in subsequent periods through the innovation factor $\gamma(B)$ where $\gamma_B < 0$ and $\gamma_{BB} > 0$.¹ In period 1, $\gamma(B)$ equals one. Only by investing in process innovation can the firm earn positive profits in period 2.

If the firm seeks to diversify into the production of the unregulated product y in period 1, it must obtain the approval of the regulator who may simultaneously lower the regulated revenue requirement and the regulated price in this period. The regulator does this by shifting responsibility for some portion, θ between 0 and 1, of the fixed cost of producing x to the unregulated market. The value of θ is independent of the amount of y produced and remains constant throughout the three periods at the level set by the regulator in period 1. The value of θ designated by the regulator reflects his attitudes about how the benefits of economies of scope should be split between the regulated and unregulated markets. (The regulator could specify different values of θ for different lines of business that the firm might choose to enter.) The higher the value of θ , the lower the portion of joint cost that is included in the regulated revenue requirement and, *ceteris paribus*, the more the consumers of the regulated product benefit from scope economies.²

Prior to making its decision about diversification, the firm learns the value of θ set by the regulator. If the firm decides to diversify and θ exceeds zero, the price of the regulated product is reduced in period 1. The profit-maximization problem for the diversified firm is³

$$(2) \quad \text{Max}_B P_y y_1 - \theta F - C(y_1) - B \\ + \delta [\text{Rev}(x_2^d) + P_y y_2 - \gamma(B) \times [F + C(x_2^d)] \\ - C(y_2)] + \delta^2 [P_y y_3 - C(y_3) - \theta \gamma(B) F]$$

where $C(y)$ is the incremental cost of producing product y in addition to product x .

It is assumed that the unregulated market is perfectly competitive with a price of P_y , therefore, the firm sells the same quantity of y , that which equates price to marginal cost, in each time period.

II. Investment in R&D

According to my model, the profit-maximizing level of R&D expenditures for the undiversified firm falls below the socially optimal level (given x_1) because the firm does not retain the higher profits associated with lower third-period costs. The diversified firm also selects a suboptimal level of investment in B ; however, it selects a higher level of B than the undiversified firm if θ is greater than zero. (Derivations of these and all further results are found in my 1990 paper.)

¹In my specification of the model, I implicitly assume that there are economies of scope between production of x and R&D to reduce the cost of producing x , when having this R&D performed by the regulated firm's suppliers is not considered. This assumption is justified if familiarity with day-to-day operations in the production of x affect the parameters of the γ function.

²Given the technology and the regulatory rules that I specify and the assumed ability of the regulator to attribute costs accurately to either a common cost category or to one of the two products, reducing the regulated product's share of common costs (increasing θ) will lower regulated price and increase output. Us-

ing an Averch-Johnson model of the regulated firm, Michael Crew and Keith Crocker (1990) find that if the regulatory constraint is binding after diversification, more of the regulated product is produced after diversification only if all production costs including those for the unregulated product are included in the revenue requirement.

³I am implicitly assuming that the firm does not engage in any wasteful spending given the chosen technology. Ronald Braeutigam and John Panzar (1989) show that if the regulator knows the breakdown of production costs between joint and attributable costs, then the firm has no incentive to make wasteful expenditures in any cost category.

Three factors contribute to the diversified firm's enhanced incentive to undertake R&D: 1) the one-period regulatory lag; 2) the initial increase in demand for the regulated product in periods 1 and 2 spurred by the reduction in the regulated revenue requirement; and 3) the opportunity to keep some of the gains from R&D in the form of higher profits in the unregulated market in period 3. The last two factors both require that θ be greater than zero. Just as increasing the length of the regulatory lag increases a firm's incentive to invest in R&D (Elizabeth Bailey, 1974), increasing the value of θ also increases the firm's preferred level of B . As θ rises, first- and second-period output of the regulated product increase and the share of the reduction in common costs that is lost to consumers in the form of a lower third-period price falls, both serving to increase investment in R&D.

The firm's decision to diversify or not depends on the value of θ set by the regulator. If θ equals zero, the firm chooses to diversify since the difference between diversified and undiversified profits is positive. However, as the value of θ increases, the profit gains from diversification fall. I designate as θ^* the value of θ at which diversified profit equals undiversified profit. The value of θ^* varies positively with the degree of economies of scope between x and y .⁴ If the value of θ set by the regulator exceeds θ^* , the firm chooses not to diversify and the attending higher gains in production efficiency do not occur.

III. Welfare Effects

A welfare analysis of this increased innovative activity with diversification suggests

that there may be, previously unrecognized dynamic gains in consumers' surplus when regulated firms diversify into unregulated industries. If the cross elasticity of demand between x and y is zero and the firm continues to charge the competitive price for y after diversification and innovation, then consumers' surplus in the market for y is unchanged. In the market for x , consumers' surplus increases in period 3 in the undiversified case as price is lowered to reflect cost reductions due to R&D investments made in period 1. If the firm diversifies and θ exceeds zero, consumers' surplus in the regulated market is higher in periods 1 and 2 due to the initial price reduction. The price of the regulated product falls again in period 3 at the time of the rate review. In each of the three periods, consumers' surplus under diversification exceeds the level attained when the firm remains undiversified because the average revenue requirement is lower in all three periods under diversification.

IV. Diversification and Specialized R&D

While diversification of regulated firms may provide an incentive for potentially welfare-improving innovations, diversification also opens the door for various forms of cross subsidy and other inefficient behaviors on the part of the regulated firm. The opportunities to profit from inefficient behavior increase as the quality of the regulator's information decreases (Timothy Brennan, 1990). However, even a regulator like the one in this model, who is able to monitor and categorize production costs perfectly, is not immune to inefficient technology choice on the part of the firm.⁵ Our regulator also is unable to prevent inefficient allocation of R&D expenditures be-

⁴While scope economies do provide a legal or political justification for regulators to disallow some portion of fixed costs, economies of scope are not necessary for the regulatory policy of disallowing fixed costs to lead to increases in R&D investment. However, *ceteris paribus*, the greater the scope economies between the existing regulated product and any unregulated product, the more attractive is that unregulated product as an additional line of business to the regulated firm.

⁵See Kenneth Baseman (1981), Brennan, and Braeutigam-Panzar. I do not address this problem of inefficient initial technology choice in the model and the regulatory policy is not designed to alleviate this problem specifically. However, if the regulator sets a high value of θ , the firm's incentive to shift to a high fixed-cost technology in order to lower marginal cost in unregulated markets is reduced considerably.

tween the reduction of joint and attributable costs.

To analyze the implications of differentiated R&D investments for the earlier results, I modify the model of R&D investment decisions to allow the firm to disaggregate its innovative effort between reducing common costs and reducing attributable costs as found in Braeutigam and Panzar. I use this model to compare R&D investment of diversified and undiversified firms and the associated reductions in the regulated price across all three time periods. I find that the diversified firm invests more than the undiversified firm both in R&D to reduce cost attributable to x and in joint-cost-reducing R&D. However, the diversified firm is biased vis-à-vis the undiversified firm toward joint-cost-reducing R&D and this bias reduces the magnitude of economies of scope over time. Despite this additional distortion in innovative effort, consumers' surplus is higher in all three periods when the regulated firm diversifies.

V. Conclusions

I have derived a model in which the regulator can induce the diversified firm to invest in R&D by making a reduction in the portion of fixed costs included in the regulated revenue requirement a precondition for allowing the firm to diversify. Under this rule the firm diversifies if the designated value of the fixed cost sharing parameter, θ , is below the firm's breakeven level, θ^* . The diversified firm picks a higher level of R&D investment and consumers of the regulated product are better off if the regulated firm diversifies because of the lower average revenue requirement. The policy of sharing fixed costs leads to higher consumers' welfare gains over time, even if the diversified firm's R&D investment is biased toward fixed-cost-reducing R&D.

The assumptions underlying this analysis are that the regulator is able to distinguish joint costs and attributable costs and to assess accurately *ex post* the effects of R&D investment on the firm's cost function. If the regulator is unable to categorize costs or to observe these cost reductions, then the

negative effects of cross subsidy or distorted technology choice or both may outweigh any long-run benefits to consumers from reduced costs. Two other important assumptions in this analysis are that the market for the unregulated product y is perfectly competitive and that the demand functions for x and y are independent. Violation of either assumption could increase the firm's incentives to cross subsidize and result in lower or negative net gains in consumers' welfare from increased innovation.

Diversification into an unregulated line of business could also increase the number of and the usefulness of new products and services that might be developed by the diversified regulated producer, potentially increasing the level of consumers' welfare over time. For example, allowing a regulated service producer to manufacture the capital equipment used to produce the service could bring a more extensive knowledge of the marketplace for the existing and for potential new services to the R&D process for the equipment manufacturer.⁶ Further research is required to explore these and other possible links between diversification and improved research performance by regulated firms before making significant changes to regulatory policies and curtailing line-of-business restrictions.

⁶Alfred Kahn (1990) uses this argument to justify allowing the regional Bell Operating Companies to manufacture communications equipment.

REFERENCES

- Bailey, Elizabeth, "Innovation and Regulation," *Journal of Public Economics*, August 1974, 3, 285-95.
- Baseman, Kenneth, "Open Entry and Cross-Subsidization in Regulated Markets," in Gary Fromm, ed., *Studies in Public Regulation*, Cambridge: MIT Press, 1981.
- Berg, Sanford and Tschirhart, John, *Natural Monopoly Regulation Principles and Practice*, Cambridge: Cambridge University Press, 1988.

- Braeutigam, Ronald and Panzar, John, "Diversification Incentives Under 'Price-Based' and 'Cost-Based' Regulation," *Rand Journal of Economics*, Autumn 1989, 20, 373-91.
- Brennan, Timothy J., "Cross-Subsidization and Cost Misallocation by Regulated Monopolists," *Journal of Regulatory Economics*, March 1990, 2, 37-51.
- Crew, Michael and Crocker, Keith, "Diversification and Regulated Monopoly," in Michael Crew, ed., *Competition and the Regulation of Utilities* Boston: Kluwer, 1990.
- Kahn, Alfred, "Telecommunications, Competitiveness and Economic Development—What Makes Us Competitive?," *Public Utilities Fortnightly*, September 13, 1990.
- Palmer, Karen, "Diversification by Regulated Monopolies and Incentives for Cost-Reducing R&D," Resources for the Future working paper, 1990.

DIFFUSION OF DEVELOPMENT[†]

Diffusion of Development: Post-World War II Convergence Among Advanced Industrial Nations

By RICHARD R. NELSON*

Over the past half-dozen years, a significant literature has grown up on the phenomenon of "convergence," the erosion of the striking U.S. lead in productivity, incomes, and technology across a wide front, that clearly existed during the quarter-century after World War II. It was Moses Abramovitz' masterly article (1986) that (arguably) started the recent flow of research and writing, and his is (arguably) still the most sophisticated statement of the phenomenon, and articulation of a broad theory to explain it.

Abramovitz posited that the presence of a nation (or a firm?) employing significantly superior technology provides others with a target to emulate, an opportunity for rapid growth, and a context where one might expect convergence. However, he also observed that, even in the post-1960s, not all nations were "converging," and employed the term "social capabilities" to designate the requirements to take advantage of the opportunity. He mentioned an effective education system and appropriate institutional structures, among other things, as components of "capabilities." Abramovitz also noted that there seemed to be something special about the post-World War II period, or the interwar period, or both, because the U.S. lead in productivity and income dated from before World War I, and yet there is little evidence of convergence until the 1960's, even among nations that, presumably, had ample capabilities, in the sense of

an educated work force, modern financial institutions, etc.

I presume that the other two papers in this session will deal with the mixed experience of the countries that were underdeveloped in 1950, and the generally bad experience of the socialist economics which, as it turned out, led to the abandonment of that kind of economic system.

I. How Did the United States Get its Lead?¹

This paper is about the convergence that has occurred among the major industrialized nations. For the most part, the convergence literature takes the U.S. lead after World War II as "given," and pays no attention to exactly what that lead was, and where it came from. My proposal here is that understanding of the sources of the U.S. lead enables one to understand better why there was failure of convergence among the advanced industrial nations prior to World War II, why convergence has occurred so rapidly since 1960, but among only a limited set of countries. My basic argument is that there were two distinguishable components of postwar U.S. technological dominance. One was the lead in mass production industries, and this was of long standing. The other was in "high-tech" and this was new. Each stemmed from different sources. And they eroded for conceptually separate, if institutionally connected, reasons.

While Great Britain was the pioneer and dominant technological power of the first

[†]*Discussants:* Paul David, Stanford University; Donald Harris, Stanford University; Henry Rosovsky, Harvard University.

*Columbia University, New York, NY 10027.

¹The basic argument developed here is presented in more detail in my 1990 article. Gavin Wright (1990) has developed an analysis of the "early" U.S. lead that is strongly complementary to mine.

Industrial Revolution, in a variety of areas the Americans caught on quickly. A number of economic and technological historians have documented the early rise of mechanized production of fire arms in the United States, and the subsequent spread of this technology to other industries.

During the last half of the nineteenth century, a rapidly growing mass market of relatively well-to-do households, and an expanding rail and communication network that broke down distance barriers, led in a number of industries to the rise of large companies engaged in mass production and nationwide marketing. The story has been well described and documented in Alfred Chandler's work. (See particularly his 1977 study) In many of these industries, American firms developed a clear lead in process technology or product design, or both, over British and continental firms. The U.S. firms dominated world production and trade in such diverse products as sewing machines, typewriters, matches, and refrigerated meat, for example. By the 1890's, the United States led the world in steel production, and American firms were widely recognized as the most efficient steel producers.

There were several factors that were special about the American scene that led to these developments. First, by the late nineteenth century, the United States had become the world's largest richest common market. In a context where exportation was not easy, American firms were in a much better position to exploit latent economies of scale than were European firms. Second, large-scale production tended to be capital intensive. High American wage rates, and relative scarcity of skilled labor, encouraged and supported such capital-intensive production. Third, reflecting America's resource abundance, the new technologies tended to be natural-resource-intensive, both directly and indirectly, in that American machinery was low priced relative to machinery produced in Europe, because American metals were cheaper.²

²The advantages lent to American industry by cheap natural resources have been stressed by Wright, and much earlier by Nathan Rosenberg (1976).

Economists looking at this situation through the lens provided by standard price theory would be inclined to propose that market conditions induced American firms to produce in a different manner than the European ones. This does not imply that Americans had command over technology that the European could not have used had conditions there been different. However, in fact, effective command of the technologies in question involved significant elements of learning by doing and using on the part of engineers, mechanics, craftspersons, and managers. Adoption of these technologies by European firms, had they wanted to do so, would have been a very costly and time-consuming business.

The American lead in mass production industries was preserved over the interwar period in large part because economic conditions in the United States continued to warrant larger-scale and more capital-intensive production than in Europe. Barriers to intercountry trade were high during this era. The American market was significantly larger than the market faced by national European firms. American real wage rates were higher, and more capital-intensive production therefore warranted.

In the early years after World War II, Americans continued to be dominant in these large-scale and capital-intensive industries. European engineers and managers clearly recognized American superiority, and that it was of long standing.

I have been arguing that the American post-World War II mass production industries had been around for a long time. In contrast, the American postwar lead in the new high-tech industries of the postwar era (jet aircraft, semiconductors, computers, telecommunications, pharmaceuticals) was a new phenomenon.

By the late nineteenth century, technological advance in the chemical and electrical industries began to take place in industrial laboratories dedicated to invention, and staffed by university-trained scientists and engineers. It was German firms, not British or American, that took the lead in the new industries associated with organic chemistry. In the new electrical equipment indus-

tries, American firms were in the forefront of some fields, but in most had no significant lead over German or British firms, and, in a number of areas, European firms were ahead.

On the eve of World War II, the major American chemical companies had more or less caught up with the Germans, and the major American electrical equipment companies were world class, and doing well in rapidly growing fields connected with radio and electronics. American universities were quite respectable in a variety of the sciences. But there is no evidence of American leadership in these high-tech fields that was on a par with American leadership in mass-production technologies.

However, it is apparent the Americans did take the leadership in the above-named fields after World War II. The Europeans recognized this, and worried about it. What lay behind this new American lead?

The answer, I would argue, lies in massive investments in research and development (R&D) that American firms, universities, and the American government began to make after the war. By the early 1960's, R&D as a fraction of GNP was roughly double in the United States what it was in Europe. This vast investment in R&D was supported and made possible by significant increases in the fraction of young Americans going on to higher education and into science and engineering.

A small but important fraction of this greatly expanded pool of scientific and engineering talent was employed in the U.S. university research system. Support for this massive increase in university research was largely provided through the government.

However, the lion's share of the increased investment in R&D occurred in American industry. Part of the increase in corporate R&D was the result of major increases in private R&D funding, based on optimistic beliefs about the profitability of such investments which, by and large, turned out to be well-founded. Part was due to massive DOD and, later, NASA investments in the system. As it turned out, Department of Defense R&D expenditure and procurement was an extremely important factor behind the rise

of U.S. industrial predominance in commercial jet aircrafts, semiconductors, and computers.³

While commentators at the time treated "spillover" from military R&D as if it were a natural kind of event, I propose here that the circumstances in the 1950's and 1960's were somewhat unusual. It turned out that the kinds of weapons the armed services wanted then involved technologies that soon had major civilian use. During the nineteenth century, army demands for guns with interchangeable parts put in place a machine-tool technology that later had widespread civilian use. However, it is arguable that, in the normal course of events, the flow is more from technology developed for civilian purposes to military needs, than the other way around, and when the military goes for a technology not available in the civilian sector, later spillover is quite limited.

By the mid-1960's, the new American lead in high-tech industries, as the old lead in mass production industries, was widely taken as a fact of life, a source of pride for Americans and of concern for Europeans, but presumably a fact not easy to change. Ironically, as we now know, by then, American dominance was fast eroding. It was shrinking both in the areas of longstanding American preeminence (mass-production industries) and in the new high-tech fields that the United States seized after the second World War.

II. Convergence

I propose that this account of the nature and origin of the U.S. post-World War II lead, or rather the twin leads, helps one to understand the subsequent convergence phenomenon and its timing. It suggests that not just one, but several, different things were going on behind the scenes.

One central thing was the opening of world trade in manufactured goods, and in

³See the chapters by David Mowery and Nathan Rosenberg, Richard Levin, and Barbara Katz and Almarin Phillips, in my book (1982).

raw materials. With that, the world became "a common market," and the American advantage in scale-and resource-intensive production, that it had had for almost a century, eroded. With national boundaries no longer such an obstacle, it made sense for European industrialists to rebuild their production capacity along American lines. That they did, and as their productivity and efficiency jumped, European and Japanese firms became competitive. Their sharply rising exports both reflect this, and were a precondition for successful adoption of scale-intensive technologies in the first place. And real wages rose with productivity thus justifying and supporting higher capital intensity.

The erosion of American dominance in high-tech industries is also a story of opening trade diminishing early American scale advantages, and involves other elements. Up until the mid-1960's, the American military market for semi-conductors, and computers, that was basically reserved for American firms, counted for a large share of total demand. By the mid-1960's, however, the civilian market for these products had outstripped the military in scale, and this market was open to foreign competition.

At the same time, technological communities were becoming increasingly transnational. Partly this was part and parcel of the internationalization of business that was occurring. Partly it was the result of the greater ease of communication and travel that marked the period. As a result, the general elements of technological knowledge became widely accessible to trained scientists and engineers, whatever their nationality.

However, development of operational capabilities in the high-tech industries required massive investments—in R&D and the training of scientists and engineers needed to do R&D, as well as in plant and equipment. By the late 1960's, the major European nations and Japan were making these investments. Scientists and engineers as a fraction of the work force, and R&D as a fraction of GNP, rose to approach American levels.

At the same time, a considerable amount of scattered evidence suggests strongly that the large American military R&D efforts, that I have argued was a plus in the 1950's and early 1960's, began to be a minus or, at least, not a help. The rise of civilian demand for semiconductors and computers, and for commercial jet aircraft, stimulated very large-scale privately financed R&D efforts to meet these demands. At the same time, military R&D increasingly was being focused on special attributes particularly valued by the military. By the 1970's, it is arguable that there was much more "spillover" from civilian R&D to the military, than the other way. Yet the United States continued to spend proportionally far more of its R&D on military projects than did the other industrial nations. By the 1980's, Germany, Japan, and several other nations were spending a larger fraction of their GNP on civilian R&D than was the United States.

Let me summarize. The U.S. technological lead after World War II had two different components: an old part in mass-production industries stemming from the fact that for over a half-century, the United States was the world's largest common market; and a new part in high-tech stemming from massive investments in R&D and training of scientists and engineers. The twin leads eroded as the world became a common market, and nations with the requisite "social capabilities" moved to catch up, and as those same nations invested in the special social capabilities needed to compete in high-tech R&D and higher education.

This is a much more elaborate and complex explanation of the convergence phenomenon than that contained in many of the recent writings concerned with it. Several of these portray convergence as something that is more or less automatic, and thus cannot come to grips with why convergence was so much more rapid in post-World War II than before, and why there are only a few members of the convergence club. In my view, Abramovitz' treatment of the matter, which stresses social capabilities, and, also, hindrances to European industry dur-

ing the interwar period, remains the most sophisticated and illuminating general discussion of the matter. I consider my treatment here an elaboration and extension of his.

REFERENCES

- Abramovitz, M., "Catching Up, Forging Ahead, and Falling Behind," *Journal of Economic History*, June 1986, 46, 385-406.
- Chandler, A., *The Visible Hand: The Managerial Revolution in American Business*, Cambridge: Harvard University Press, 1977.
- Nelson, R., "U.S. Technological Leadership: Where Did it Come From and Where Did it Go?," *Research Policy*, Summer 1990, 19, 117-32.
- _____, *Government and Technical Progress: A Cross Industry Analysis*, New York: Pergamon Press, 1982.
- Rosenberg, N., *Perspectives on Technology*, Cambridge: Cambridge University Press, 1976.
- Wright, G., "The Origins of American Industrial Success, 1879-1940," *American Economic Review*, September 1990, 80, 651-68.

Diffusion of Development: The Soviet Union

By MARSHALL I. GOLDMAN*

Younger economists often find it difficult to believe that, at one time, the Soviet approach to economic development, or at least the sanitized version of it, was often held out as a worthy model for developing countries. Even by most alternative calculations of Soviet economic growth, the Soviet Union seemed to be making great strides. The Soviet Union showed that they had mastered advanced technology by being the first to send a man into space, while on the ground they reported record harvests and some of the world's fastest growth rates in steel production and in the output of other basic raw materials. The Soviets seemed to have transformed Western development techniques—highlighting some aspects and abandoning others. Until the mid-1960's at least, the Soviet Union seemed to have mastered the secret of economic development and there seemed to be some basis to Nikita Khrushchev's claim that, by 1970 or 1980, the Soviet Union would be able to overtake the United States and move on to the attainment of full communism.

If the Soviet economy was doing so well, why is it that suddenly it seems to be collapsing? Nor is it just a simple slowdown. Many Soviet economic institutions have simply disintegrated. What is happening now resembles the breakdown in some Third World countries, or parts of Europe and the United States in the early 1930's. There is no simple explanation for what has happened. But, in part, some of the Soviet Union's current problems are a consequence of the approach it took to development in the 1920's and 1930's. For a time, Soviet leaders assumed that they could suspend or ignore many of the governing laws of economics, keeping what suited them and

rejecting what did not. They are now discovering that it is dangerous to mess with macro- and microeconomics.

I

Stalin was determined to accelerate his country's economic growth. He was worried about a potential military threat from Germany, and he realized that he had few friends and even fewer prospects for revolution in the capitalist world. Thus he concluded that he had to move fast and, as a consequence, he decided he could not rely on traditional and time-consuming economic methods to jolt his essentially agrarian society into an advanced industrial powerhouse.

Without spelling out exactly what it was that he was doing (he probably was not fully aware of the theoretical consequences of what he was doing since they were much less fully understood at the time), in effect, Stalin declared macroeconomics null and void. There would still be taxes, government expenditures, borrowed money, and interest rates, but there was no thought that these procedures would be used to fine-tune the economy, or to increase or decrease the size of the multiplier. At most, these were necessary mechanisms for the running and financing of the government and specific industries.

Microeconomics was set aside in much the same way. In a sharp departure from other economies at the time, Stalin decreed the nationalization and collectivization of all the means of production. All factories, shops, and farms became state entities. As such, state authorities assumed the right to issue orders about the most elementary activities. Even the decision to import or export, as well as control over foreign currency, became a monopoly of the Ministry of Foreign Trade.

Since they owned and determined the behavior of almost all the country's means

*The Kathryn W. Davis Professor of Soviet Economics at Wellesley College and Associate Director, Russian Research Center, Harvard University, Cambridge, MA 021138.

of production, finding a better equilibrium of supply and demand was not something Soviet leaders worried about. Besides, most of those put in charge of the economy were engineers who concerned themselves more with increasing production, especially of heavy industrial products. To bring about the accelerated growth that Stalin wanted, the country's engineers and planners came up with the idea of the yearly and 5-year plans. Shortages were essentially treated as short-term phenomena, that sooner or later were sure to be eliminated through improved planning procedures and accelerated plans. There was seldom any concern that, at existing prices, there were too many customers and not enough supplies. These temporary glitches would ultimately disappear and production would increase. That was the main priority, and increases in production became the measure for judging factory manager as well as national economic performance. In effect, an administered system was put in place in which planners, administrators, Gosplan, Gossnab, and the various industrial ministers, including the Ministry of Foreign Trade, set industrial output targets and allocated supplies of capital and labor. This superseded what in other societies was determined by individual decision makers guided only indirectly by the government's monetary and fiscal policies.

Inevitably, the Soviet system gave rise to enormous disproportions, some of which had serious consequences. The full measure of what was happening, however, was masked by the unrestrained use of the country's abundant stock of raw materials and labor. This practice was institutionalized by the deliberate decision to set low prices on raw materials. The low raw material prices were combined with an incentive system called "val." The val system rewarded managers for increasing the gross ruble value of their output compared to what had been reported the year before. Since gross ruble output was measured simply by adding up the costs of inputs, managers were in effect induced to use as many inputs as possible. As a result, Soviet manufactured products became heavy, and energy and raw material became wasteful. This was exemplified in the way the Soviet Union built airplanes.

They were heavy and required extra powerful engines. This made Soviet planes unsalable in the capitalist world because commercial airlines were unwilling to pay the extra money needed to fuel those engines.

None of this was particularly disturbing to Soviet authorities. There were, after all, abundant reserves of raw materials that could be tapped by simply assigning more of the Soviet Union's abundant labor reserves to Siberia. Even some American economists supported such a utilization of resources. Given what at the time seemed to be the Soviet Union's abundant oil reserves, there seemed to be nothing wrong with the Soviet Union increasing its exports of petroleum in order to pay for increased imports of grain. (See Ed Hewitt, 1983; John Vanous, 1982.)

The only flaw in this reasoning is that Soviet energy supplies are limited and as more readily accessible supplies are used up, it becomes more and more costly, even by Soviet standards, to find additional reserves. Moreover, quantity of output increasingly became a less important criteria for determining economic success. With the growth of high technology and miniaturization, it was quality that counted. It was time to call off the race to build the world's largest microchip.

II

The Soviet administrative planning system was ill-suited to deal with the shift in emphasis from quantity to quality. Procedures designed to take advantage of the Soviet strengths under the old system seemed more and more inappropriate, if not counterproductive. This is well-illustrated by a look at what happened when the Soviet Union began to pay more attention to foreign trade.

Under the old regime, foreign trade was the monopoly of the Ministry of Foreign Trade and its foreign trade organizations (FTOs). This, it was thought, would give the Soviet Union an economic advantage in dealing with the anarchy of the marketplace. As a solitary buyer for the whole country, the Ministry of Foreign Trade was able to play off one contending capitalist seller against another in order to obtain the

lowest price possible. This ability to manipulate made possible the "great grain robbery" of 1972.

The monopoly also fit in well with the Soviet Union's political concerns. For example, for security reasons, Soviet authorities insisted that contact between Soviet and foreign trading partners be strictly controlled. In the vast majority of instances, that meant that foreign buyers and sellers usually had no contact with their Soviet suppliers or buyers. All transactions were transmitted through the filter of the FTO. This precluded feedback or after-sale service.

In foreign trade, as elsewhere in the economy, the impact of macro- and micro-economic considerations was severely circumscribed. While Soviet officials did seek to insure some level of equilibrium between hard currency imports and exports, they did not let the market determine access to hard currency reserves. Exporters, particularly the ministries of oil and gas, the source of 60–70 percent of the Soviet Union's hard currency earnings, received almost none of the proceeds. Hard currency or *valuta* was instead allocated on an arbitrary basis. Political considerations seemed to be as important as economic criteria. For example, the chemical industry, whose Minister Leonid Kostandov seemed particularly well connected, was consistently allocated 25 percent of the hard currency set aside for machinery imports, even though the chemical industry accounted for only 8 percent of the Soviet Union's total industrial investment. (See Philip Hanson, 1981.)

Since there was no need to repay that hard currency (which was fortunate since the earnings from the anticipated surge in petrochemical exports never seemed to materialize), then *valuta* was in effect a free good. Thus no one had to worry about the suitability of scale or payback periods, or feasibility studies. This contributed to gigantomania or the large-typewriter-carriage syndrome. Political clout in an office was easily measured by the size of its typewriter carriage. Important offices had big carriages—unimportant offices had small carriages. It was immaterial whether or not those ex-

tra large carriages were ever used—some day they might be and that was enough. Machinery specifications were determined in the same way.

Politics even played a role in plant location. Factory locations would occasionally be determined by the need to repay a favor to a local official, rather than to access to the necessary raw materials. Based on some interviews with Western exporters, I was told that an American-built milk plant, for example, was located in an area where there was a shortage of cows. European and American chemical company executives reported that they were told to install a petro-chemical plant in Tomsk and Omsk, even though the needed raw materials were not readily available in those regions.

Political considerations also accounted for the persistent refusal of ministries to cooperate with one another. They gave preference, instead, to units within their vertical jurisdiction. Thus according to Western exporters, individual ministries normally refused to share information with one another about Western imports, and similarly refused to join together to set up such things as a parts center or a wholesale operation. In the administrative planned system, that might actually complicate rather than simplify operations.

Unlike South Korea, where there was also government intervention and where prices were also insulated from some market forces, Soviet managers were not made to compete in world markets. Thus there was no reason to seek to enhance product quality or competitiveness. In effect, Soviet managers were sheltered from the competitive rigors of foreign trade, that could have served as a stimulative force in the same way as Janos Kornai's hard budget constraints.

III

By the time Mikhail Gorbachev became General Secretary of the Communist Party, there was fairly widespread agreement that the existing system had not helped the Soviet Union compete with the West. It was not just that the quality gap seemed to be

growing, the Soviet Union also seemed to be falling behind quantitatively. Thus Gorbachev set out to reduce, if not abolish, many of the features of the centrally planned system. Ministries were abolished and reorganized, and ministries and chairmen of Gosplan were continuously replaced in an effort to find someone who could remedy the Soviet Union's problems.

Unfortunately for Gorbachev, he succeeded in destroying portions of the old order before he could replace their functions with processes operating under the new order. As a result, he found himself having to deal with some of the worst aspects of both communism and capitalism. He began to unleash macro and micro forces before he could put in place the restraints that evolved in capitalism to temper their impact.

The best illustration is the macro consequence of inflation. It turns out that the Soviet Union had incurred an annual budget deficit for at least a decade before Gorbachev came to power. But, like a low-grade infection, it was not considered to be anything worth worrying about, particularly in an environment where macroeconomics was considered to be nonoperational. Thus Gorbachev paid little heed to the effect of increasing expenditures on machine tools and decreasing revenues because of the curb on the sale of vodka. The 18 billion ruble budget deficit in 1985, Gorbachev's first year, differed little from the deficits of preceding years. In 1986, however, there was almost a threefold increase in the deficit, and the problem was no longer trivial. By 1989, the deficit exceeded 100 billion rubles, equivalent to about 10 percent of the Soviet Union's GNP. By 1990, no one seemed to know how big the deficit was any longer. By some estimates, it may have reached 200 billion rubles.

Belatedly, Gorbachev came to appreciate the significance of what had happened, but by then it was too late. In his own words, he acknowledged that in October 1990, "we lost control over the financial situation in the country. That was our most serious mistake in the three years of perestroika." And 5 years too late, he went on to note,

"achieving a balanced budget today is the number one task and the most important one" (both citations found in *The New York Times*, October 20, 1990).

The reason for Gorbachev's concern was that, by 1987, the deficits that were financed in large part by the issuance of money, gave rise to an increasingly large inflation. Again, no precise figures are available but, by 1989, some estimates indicated that inflation was at the level of 20 percent a year, and that there was a growing fear of hyperinflation. Inflation, in turn, led to hoarding as consumers tried to protect themselves from the fall in the value of the ruble. The result was a collapse of the distribution system. As Vasily Selyunin, an acute economic observer, put it, "during the second half of last year [1989], trade began to fall to pieces before our eyes" (*Sotsialisticheskaia industriia*, April 6, 1989).

The irony is that just when the Soviet Union seemed ready to open the institutional doors to market processes, the Soviet Union found that it first must pay its bills for all those years of denying the central role of economics. The Soviet Union today has to deal with inflation, shortages, and economic separatism. The decision to end the monopoly of the Ministry of Foreign Trade by abolishing the Ministry has caused similar problems. The expectation was that with no monopoly control, foreign trade might some day become more responsive to Soviet consumer wants. In the meantime, however, that lack of central control has resulted in a record trade deficit, and, by 1989, the Soviet Union found it could not pay the bills owed to foreign suppliers.

It was not only the economic, but the political, bills of the past that have begun to frustrate Gorbachev's reform efforts. After several months of indecision and policy reversals, Gorbachev finally decided to encourage the formation of private and cooperative businesses and family farms. The move to private business was slow at first, but ultimately attracted some venturesome entrepreneurs. But the indecision and continuing controls on such activity serves to prevent an increase in supply large enough to bring down prices. The expansion of the

supplies of flowers in Soviet cities proves that such an increase is possible, but more typical is the reluctance of the peasants, especially in the Slavic parts of the country, to pick up Gorbachev's offer to set up their own family farms. Remembering what happened to the kulaks in the 1930s, as well as how Khrushchev simply reduced the size of their plots and the number of their cows in 1960, the peasants decided that the higher return that they might obtain from an investment in land and livestock was not worth the risk that they would lose it all after another shift in the political pendulum. Thus, as of July 1, 1990, after almost 3 years of trying, there were only 30,000 family and private farms in the entire Soviet Union. Over one-half of these farms were in the republic of Georgia and one-third were located in Latvia, Lithuania, and Estonia. Meanwhile there were only 900 of these farms in the Russian Republic and a mere 12 farms in the Ukraine (*Ekonomika i zhizn'*; November 1990).

To an economist, the effect of all these upheavals is that the time horizon of the average Soviet citizen has been cut back sharply. Few are willing to invest in the future, which of course makes it all the more difficult to cope with the present. The minute any one in the Soviet Union smells money, especially foreign *valuta*, they make a grab for it. Thus even joint ventures like Combustion Engineering and McDonalds have fallen victim to this pressure. Instead of taking such efforts and turning them into showcases, that would presumably attract similar investments from outside the Soviet Union, the Soviets instead milk them immediately. Thus, in November 1987, Combustion Engineering invested \$8 million in a joint venture called Applied Engineering Systems. This led to an immediate increase in the yield of some Soviet oil refineries and profits in dollars for the American partner. However, after a while, the Soviet partner began to disregard the need for financial controls. Fancy cars became more important than basic engineering, and, by November 1990, the American partner found it necessary to withdraw its staff from the Soviet Union.

McDonalds encountered similar problems. Rather than wait for profits to accrue from a profitable operation, the Soviet partner decided instead to take its money out at the front end of the operation. In less than 10 months time, the Soviet partners (the Moscow City Council), decided instead to raise the joint venture's office rent tenfold and the price of the meat that the joint venture used by sixfold. Since McDonalds was not selling for dollars, that meant that the local authorities were not even prepared to wait for *valuta* to be earned. McDonalds in turn found it necessary to double prices simply to pay its ruble bills. In the meantime, its Moscow partner was able to raise the yield far above that needed to stay ahead of inflation. The message for the hesitant investors from the West is hesitate some more.

IV

Because of distortions and distrust brought on by Stalinist central planning, the belated adoption of traditional economic market stimuli had a perverted rather than a positive impact. Instead of facilitating healthy economic growth and constructive responses, the tendency has been to put a premium on instantaneous profit maximization. Like the corn hog cycle cobweb that explodes rather than converges, the diffusion of normal economic processes in a severely repressed society like the Soviet Union works to exaggerate rather than dampen economic distortions.

He is usually quoted in conjunction with political change, but Alexis de Tocqueville's comments that "the perilous moment for a bad government is when that government tries to mend its ways," seems to be just as apt when the government decides to acknowledge the reality of macro- and micro-economics.

REFERENCES

Hanson, Philip, *Trade and Technology in Soviet-Western Relations*, New York:

- Columbia University Press, 1981.
- Hewitt, Ed, "Tinkering with the Soviet Economy," *The New York Times Book Review*, July 10, 1983, pp. 11, 28.
- Vanous, John, *Comparative Advantage in Soviet Grain and Energy Trade*, Philadelphia: Wharton Econometric Forecasting Associates, 1982.
- Ekonomika i zhizn'*, November 1990, No. 46, p. 13.
- New York Times*, October 20, 1990, p. 6.
- Sotsialisticheskaia industriia*, April 6, 1989, p. 1.

Diffusion of Development: The Late-Industrializing Model and Greater East Asia

By ALICE H. AMSDEN*

A striking feature of late industrialization is its regional bias. East Asia, both north and south, has encompassed many of the fastest industrializing countries, including Japan, South Korea, Taiwan, Malaysia, Indonesia, and Thailand (Singapore and Hong Kong, although prospering, are distinct because they never transformed themselves from an agrarian base, the usual meaning of "industrialization"). Between 1960 and 1988, industry grew annually by 10.6 percent in East Asia, but by only 6.3 percent in all developing regions (World Bank, 1990).

East Asian growth has not been uniform; the Philippine's performance has lagged. But there are no analytical shortcuts to determine why East Asia generally has boomed by contrasting it with the deviant Philippine case. For example, for over a decade the Philippines promoted exports, a policy that is sometimes credited with greater East Asia's success. Between 1973-80 and 1980-86, the Philippine's ratio of manufactured exports to manufactured value-added nearly doubled, but its economy stagnated (as did the export-oriented economy of Puerto Rico) (UN ESCAP, 1990). Thus, the key to understanding East Asia's success must be sought beyond export activity (or education, in which the Philippines and Puerto Rico have excelled as well).

Equitable land distribution, a structural factor general to East Asia but not to the Philippines, probably does go far in explaining divergent intraregional (and interregional) growth rates. East Asia's equitable land distribution by world standards, a virtue in itself, also appears to underlie its equitable distribution of wages and salaries, which has helped keep social conflict and inflation in check, which, in turn, seems to

encourage high savings rates. Nevertheless, the relationship between industrialization and income distribution is more complex than the above comparison would suggest. Brazil, for example, with a highly skewed income distribution, grew exceptionally fast for over 25 years. Growth has also been rapid in East Asian countries whose income distributions are not very equitable—say, Thailand, and Indonesia. In the 1980's (or 1970's in some cases), the income share of the top quintile of the population exceeded that of the bottom quintile by a factor of 4.0 in Japan, 4.3 in Taiwan (Kuo-Ting Li, 1988), 4.9 in South Korea, 7.6 in Singapore, 11.2 in Thailand, 11.9 in Indonesia (rural income, Alan Gelb, 1988), 12.1 in Hong Kong, and 16.1 in the Philippines (the comparable figure for Mexico, Brazil, and the United States was 15.4, 27.7, and 10.7, respectively). (See UN, 1985.)

Evidently no single variable can substitute for a more general theory of late industrialization to explain divergent growth trends. At least two approaches to the formulation of such a theory are possible. One is to apply the "flying geese" approach (originally presented by Akamatsu), according to which transfers of technology, foreign investment, and trade between richer and poorer parts of a region generate development through a restructuring in the division of labor (see M. Shinohara, 1972). Another is to analyze the general model behind late industrialization and then to theorize why only some countries have been able to assimilate it.

Without question, the East Asian region has reaped positive (and some negative) externalities from overseas Chinese entrepreneurs and from Japan, the late-twentieth century's most dynamic economy. Nevertheless, the flying geese metaphor raises more questions than it answers and is no substitute for the second approach. By

*Professor of Economics, New School for Social Research, 65 Fifth Avenue, New York, NY 10003.

way of illustration, when American industry was "number 1" worldwide in the 1950's and 1960's, the flying geese pattern ought to have obtained in the Western Hemisphere, given large flows of technology, investment, and trade between North and South America. Yet neither has most of South America developed nor is it obvious that whatever development has occurred reflects regional externalities. Regional resource transfers and restructuring have to be placed in a larger analytical context to make sense, particularly in East Asia, all of whose fast-flying geese seem to be guided by a similar radar system.

I. Industrializing Through Learning

There are two potential models to explain late industrialization: an institutional one, as outlined below, and a market-oriented one (with overlap between the two). In the market model, industrialization is a matter of "getting the prices right" and specializing. Low-wage countries are supposed to develop by exporting labor-intensive commodities (ignoring raw materials), and presumably the fastest growing countries are those that most closely follow their comparative advantage. Nevertheless, factor proportions, the pillar of price theory, does not capture the dynamic of industrial change in either the eighteenth or nineteenth centuries. The first industrial revolution in Great Britain was largely driven by a series of technological changes that, for all practical purposes, are exogenously determined in the market model. Come the Second Industrial Revolution a century later, neither Germany nor the United States industrialized by competing against Britain on the basis of low wages. In fact, German and American wages tended to exceed Britain's. Both Germany and the United States out-competed Britain on the basis of a new wave of innovations and another phenomenon that sits uneasily in the market model—economies of scale.

In the twentieth century, the defining characteristic of industrializing "late" is the absence of new technology—even in leading enterprises. Beginning with Japan, late in-

dustrializers have initially not had the competitive asset of pioneering products and processes, which is what differentiates them from earlier industrializers (see my 1989 study). Innovators also borrow technology from their competitors, but late-industrializers are entirely dependent on "learning" in order to compete. This imperative is what gives a common development dynamic to an otherwise diverse set of late-industrializers, say, Malaysia, Mexico, Turkey, and Japan (although in many respects, Japan is unique).

In the market model, the answer to the question of how learners can export or compete against imports without pioneering technology is simple: if they follow their comparative advantage and get the prices right, then they can compete on the basis of low wages. This presumes that lower wages in the least-developed countries can triumph over higher productivity in countries one notch more industrialized. The general assumption is that production functions are everywhere identical so that the most labor-intensive commodities are indisputably the comparative advantage of the lowest-wage producers.

The empirical evidence from East Asia, however, casts doubt on this assumption. For example, a study in the 1930's to ascertain why the Japanese textile industry was bankrupting Lancashire concluded that Japan's lower wages were not responsible (see G. E. Hubbard, 1938). Wages were discounted as the critical variable in the presence of Britain's segmented labor market. Wages of young, female labor in Lancashire were not much higher than in Japan's textile mills. If wages alone mattered, Britain's textile industry might have victored. Instead, Japan's superiority was attributed, among other factors, to its more modern production facilities that, from the perspective of market theory, should have been capital-rich Britain's strong suit.

The superior infrastructure, production equipment, and management of Japanese textile companies in the 1960's made it impossible for Korean and Taiwanese textile companies to compete against them exclusively on the basis of lower wages, even after Korea and Taiwan liberalized their

economies and got the price of their exchange rate right to the point of satisfying the Bretton Woods institutions. Even in a labor-intensive sector like textiles, and even with aid-financed, modern infrastructure, the governments of South Korea and Taiwan *had* to intervene to offset Japan's higher productivity with a wide range of subsidies, far wider than those warranted to support infant *innovators* in the second industrial revolution. Once the exchange rate was devalued, subsidies were used deliberately to get prices "wrong" in order to stimulate investment and trade.

In the market model, the plight of a low-wage country that cannot compete in labor-intensive industries against the higher productivity of a higher-wage country is resolved by introducing either inward direct foreign investment from more technologically advanced countries or further exchange rate devaluations. Yet no foreign investor would invest in a low-wage country if it could make higher profits in a lower-cost, higher-productivity one. Further currency devaluations might reduce real wages and increase competitiveness, but not by much (if at all) if they raised the cost of imported wage goods (or other production inputs).

This dilemma is tautologically dismissed in price theory as a "market failure." In practice, even after devaluing, most countries struggling to industrialize without a new product or process cannot compete on the basis of low wages in a critical mass of industries, a fortiori, when their infrastructure and educational systems are relatively poor. Natural resources and handicrafts may provide a supplementary source of capital accumulation and an alternative engine of growth—as in Malaysia, Indonesia, and Thailand. Yet these countries are exceptionally rich in natural resources, and have still subsidized their non-resource-related industries to foster long-term growth.

II. Subsidy Allocation Principles

East Asia has generally been more successful at industrializing than other learners not because it has bowed deeper at the altar of free markets, but because it has operated

with a different subsidy allocation principle. This has increased productive efficiency so that the "wrong" prices have been right, and fewer subsidies have been needed to create a cost advantage. In slower-growing late-industrializing regions, subsidies have tended to be allocated according to the principle of giveaway, in what has amounted to a free-for-all. In East Asia, beginning with Japan, there has been a greater tendency for subsidies to be dispensed according to the principle of reciprocity, in exchange for concrete performance standards with respect to output, exports, and eventually, R&D.

For example, the government in South Korea disciplined its big business groups by means of price ceilings, controls on capital flight, and incentives that made diversification into new industries contingent on performing well in old ones. Thailand's Board of Investment made subsidized credit and protection from competitive imports dependent on its clients' compliance with sequential performance standards, including exports and local content targets. Taiwan's state-owned bankers have been held personally responsible for loans to business, and have carefully monitored them.

Late industrialization everywhere has involved a high degree of discipline of labor, but what distinguishes East Asia is not just its discipline of labor, but also its discipline of capital. Because a low-wage advantage cannot offset a high-productivity advantage in enough industries to spur development, government intervention is a necessary evil. The more subsidy allocation is disciplined and monitored, the faster growth.

III. Import Substitution cum Export Activity

The counterproposition, that the right market prices are necessary and sufficient to promote industrialization, implies that exporting requires no prior period of subsidization. It is enough that as a country's factor endowments change, it gains new comparative advantages, whereupon exporting can begin at once.

Nevertheless, a large share of East Asia's manufactured exports has involved a lengthy period of subsidization (during which time

incentives seem to have favored domestic sales over exports). Taiwan's and South Korea's textile exports in the 1960's arose on the basis of aid-financed incentives in the 1950's. For 25 years no foreign cars were to be seen on Korean roads and no Korean cars were to be seen on foreign roads. Most of Thailand's and Indonesia's non-resource-related manufactured exports in the mid-1980's began receiving protection in the mid-1960's.

Granted the provision of what has become standard "standby" incentives to exporters regardless of industry, the policy problem is how to insure that after an industry is subsidized for possibly lengthy periods, it achieves world-competitive levels of productivity and quality. Such micro efficiency lies at the heart of East Asia's export success.

The interconnected factors behind East Asia's micro efficiency extend from the shopfloor level to the state. Briefly, in all late-industrializing countries, the strategic focus of the firm tends to be on the shopfloor, because that is where borrowed technology is made to work. East Asia has managed the shopfloor exceptionally well, however, because the wage gap dividing managers and workers is relatively narrow, educational levels are high, performance-based bonuses are a large share of wage payments, etc. In most late-industrializing countries, the widely diversified business group is the predominant form of enterprise, sometimes large in scale (as in Japan and Korea), sometimes small in scale (as in Taiwan and Thailand). For a variety of reasons, however, such enterprises have flourished in East Asia and have diffused "best practice" management techniques to sundry industries. Finally, East Asian micro efficiency has excelled because, as argued above, the subsidy allocation process at the macroeconomic level has been relatively disciplined, and government has been able to prevent the "wrong" prices from being incorrect.

Market forces and the state have divided the labor of disciplining East Asian business. During an industry's import substitution phase, the state has typically been the disciplinarian while during its early export

phase, that role has fallen to the market. Then during an industry's "neo-import substitution phase," when subsidizing R&D and shifting into a higher-quality market segment come on the agenda, the state's dominant role resumes—as became evident by the late 1980's in Taiwan, South Korea, Singapore, and even Hong Kong.

IV. The East Asian Puzzle

An institutional model of late industrialization can be said to be taking shape conceptually. Whereas industrialization in the eighteenth and nineteenth centuries was propelled by new products and processes, late industrialization is being driven by borrowing technology or "learning." In the absence of pioneering technology, low wages even in labor-intensive sectors usually fail to provide a cost advantage at market-determined prices. Persistent problems of competitiveness even after sharp exchange rate devaluations have compelled the state to play a more active role than in the past. The absence of pioneering technology has also meant that business enterprises are less structured around a single technology than previously, and are more solicitous of shopfloor productivity and quality than firms that evolved as innovators, with a strategic focus on design and R&D.

This general model of late industrialization appears to work only under certain conditions, however, and the theoretical task ahead is understanding more about what these conditions are, at both the macro and micro levels. The foregoing discussion focused only on the former, where the East Asian evidence suggested that in subsidy-dependent industrialization, growth will be faster the greater the degree to which the subsidy allocation process is disciplined and tied to performance standards—exports possibly being the most efficient monitoring device. The disciplinarian of business activity has thus shifted over time, from simply a competitive market structure in the first industrial revolution, to Schumpeterian gales of technological change in the second industrial revolution, to an interaction of market forces and state intervention in late industrialization.

After specifying the general model, and stating the conditions under which it is likely to work, the final task is determining why such conditions are present in some countries but not in others. This task is probably the hardest, analogous analytically to pinpointing why particular companies succeed or fail. The puzzle of late industrialization has become identifying those qualities that are intrinsic to East Asia which have enabled its state to discipline business, and its diversified business groups to motivate labor, better than other regions.

Whatever role culture has played in East Asia's development, it cannot simplistically be invoked to explain its states' developmentalism. The exemplary East Asian states after the 1960's were egregious rent seekers in the 1950's—to wit, Chiang Kai-shek in Taiwan and Syngman Rhee in South Korea. The Taiwanese and South Korean states only became developmental pragmatically. Once they began not just to subsidize business but to impose performance standards on it (not least of all export targets), then growth increased. As growth increased, the state became more committed to economic development and allocated more resources to it, which increased development further. Thus, the state transformed the process of

economic development and, in turn, was transformed by it.

REFERENCES

- Amsden, Alice H., *Asia's Next Giant: South Korea and Late Industrialization*, New York: Oxford University Press, 1989.
- Gelb, Alan (and Associates), *Oil Windfalls: Blessing or Curse?*, New York: Oxford University Press, 1988.
- Hubbard, G. E., *Eastern Industrialization and Its Effect on the West*, London: Oxford University Press, 1938.
- Li, Kuo-Ting, *The Evolution of Policy Behind Taiwan's Development Success*, New Haven: Yale University Press, 1988.
- Shinohara, M., *Growth and Cycles in the Japanese Economy*, Tokyo: Institute of Economic Research, Hitotsubaki University, 1972.
- UN, *National Accounts Statistics: Compendium of Income Distribution Statistics*, New York: UN, 1985.
- UN Economic and Social Commission for Asia and the Pacific, *Restructuring the Developing Economies of Asia and the Pacific in the 1990s*, New York: UN, 1990.
- World Bank, *World Development Report 1990*, Washington: World Bank, 1990.

THE ECONOMIC IMPACT OF IMMIGRATION[†]

Immigrants in the U.S. Labor Market: 1940–80

By GEORGE J. BORJAS*

Immigration is an increasingly important component of demographic change in the United States. The significant role played by immigration in recent years sparked the development of a large literature analyzing a fundamental aspect of the immigrant experience: How do immigrants perform in the U.S. labor market?¹ A key result in this literature is that the skills of successive immigrant waves declined in the past two or three decades.

This paper continues to explore the extent and causes of the decline in immigrant skills during the postwar period. Prior to 1965, immigration to the United States was guided by the national-origins quota system. This visa allocation system awarded visas to countries based on the representation of the national origin group in the U.S. population as of 1920. The 1965 Amendments abolished the national-origins formula, thus redistributing visas across source countries, and established a system where visas are mainly given to relatives of U.S. citizens or residents.

The empirical analysis shows that a single factor, the changing national origin mix of the immigrant flow, is mainly responsible for the decline in immigrant skills.

I. Data and Descriptive Statistics

The study uses the *Public Use Samples* from four of the five decennial censuses

available since 1940. The 1970 and 1980 Censuses report the calendar year of immigration, while the 1940 and 1960 Censuses report the place of residence 5 years prior to the survey.² These four censuses allow the creation of a “recent immigrant” sample: the group of persons who arrived in the 5-year period prior to the census. Inter-censal comparisons then document the changing skills across successive cohorts. In each census, the analysis is restricted to men aged 25–64 who do not reside in group quarters, who are not self-employed, and who were employed in the civilian sector in the year prior to the census.³ Initially, the recent immigrant sample in each census will be compared with a similarly aged group of natives.

I focus the analysis on three variables: years of completed schooling, the (log) wage rate, and the (log) wage rate adjusted for differences in socioeconomic characteristics between immigrants and natives (including education, age, marital status, and metropolitan residence).⁴ Panel A of Table 1 reports the differences observed between recent immigrants and natives in each of the censuses.

Consider initially the educational attainment of the native and recent immigrant

[†]*Discussants:* Lawrence Katz, Harvard University; Stephen J. Trejo, University of California-Santa Barbara; John Abowd, Cornell University.

*Department of Economics, University of California-San Diego, La Jolla, CA 92093, and Research Associate, NBER. I am grateful to the National Science Foundation (Grant No. SES-8809281) for financial support.

¹The literature is surveyed in my 1990 book and by Michael Greenwood and John McDowell (1986).

²The 1950 Census does not provide any information on year of immigration (except for place of residence in 1949).

³The 1/100 samples of immigrants and natives contained in the 1940 and 1960 *Public Use Samples* are used in the analysis. The 1970 immigrant extract contains a 2/100 sample, while the 1980 immigrant extract contains the entire 5/100 A File. The 1970 sample of natives is a 1/1000 extract, while the 1980 sample of natives is a 1/2500 extract.

⁴To calculate the adjusted wage differentials, I estimated separate wage regressions for natives and immigrants in each census. The adjusted differential is evaluated at the sample mean of immigrants in each census.

TABLE 1—EDUCATION AND WAGES OF RECENT IMMIGRANT COHORTS IN POSTWAR PERIOD

Differences Between Recent Immigrants and Natives	1940	1960	1970	1980
A. Aged 25–64				
Education	.753	.412	-.222	-.664
Log Wage	-.031 ^a	-.128	-.160	-.299
Adj. Log Wage	-.026 ^a	-.113	-.149	-.224
B. Aged 18–24				
Education	-.231	-.285	-.835	-.313
Log Wage	.521	.331	.249	.215
Adj. Log Wage	.353	.472	.210	.128
C. Aged 25–44				
Education	.326	-.286	-.862	-1.354
Log Wage	.011 ^a	-.126	-.148	-.257
Adj. Log Wage	-.107	-.143	-.136	-.221

^aThe *t*-statistics of these differences are less than |2|. All other *t*-statistics exceed |2|.

populations. In 1940, the typical newly arrived immigrant had about .8 years more schooling than the typical native. This educational advantage decreased for subsequent waves, and by 1970, the typical newly arrived immigrant had slightly less schooling than natives. The decline in the relative schooling of immigrants accelerated during the 1970's, so that the most recent immigrants enumerated in the 1980 Census had .7 years fewer schooling than natives.

The trends in the relative immigrant wage show that recent immigrant waves have lower earnings capacities than earlier waves (see my 1985 article). Table 1 indicates that the skill decline can be observed over the entire postwar period. In 1940, the wage rate of recent immigrants was about 3.1 percent lower than that of natives. The wage differential increased to 12.8 percent by 1960, to 16.0 percent by 1970, and to 29.9 percent by 1980.

Finally, the data indicate that the deterioration in the immigrant wage cannot be explained by the relative decline in immigrant educational attainment (or by changes in other observable demographic characteristics). In 1940, the typical recent immigrant earned about 2.6 percent less than a demographically comparable native. The wage disadvantage of recent immigrants relative to comparable natives increased to 11.3 per-

cent in 1960, to 14.9 percent in 1970, and to 22.4 percent in 1980.

These intercensal comparisons implicitly assume that by differencing immigrant skills and wages from those of natives in the same calendar year, I net out the impact of the business cycle, of shifting skill prices, and of other macroeconomic fluctuations on the experiences of immigrants. It is unlikely that immigrants and natives respond equally to cyclical changes in the economy. For instance, it may be the case that immigrant wages are much more sensitive to economic downturns than those of natives. This hypothesis would explain why immigrant labor market performance lagged in 1980 (though the hypothesis would be hard-pressed to explain the 1940 data).

To ascertain the sensitivity of the results to the specification of period effects, I calculated the immigrant/native differentials using alternative reference groups. Panel B of Table 1 presents the differences between recent immigrants and young native men (aged 18–24). These two groups have one factor in common; both have just entered the U.S. labor market. If job opportunities for new labor market entrants are more sensitive to changing economic conditions, intercensal comparisons of recent immigrants that adjust for the changes experienced by young native men should provide better estimates of the secular trend.

Alternatively, one can argue that recent immigrants should be compared not to young native men, but to native men who are roughly in the same stage of the working life. A disproportionately large number of the recent immigrants in my sample are between the ages of 25 to 44 (in 1980, 85.5 percent of recent immigrants are in this age group as compared to 48.5 percent of natives). Hence an alternative base is the group of native men aged 25–44. Panel C of Table 1 reestimates the various differentials using these natives as the reference group.

Despite the major changes in the specification of period effects, the qualitative nature of the results is generally unaffected (the exception being the trend in the educational attainment of immigrants relative to that of young native men). The typical re-

cent immigrant in 1960 earned about 33.1 percent more than a young native man. The immigrant advantage over young natives declines to 24.9 percent in 1970 and to 21.5 percent in 1980. Between 1960 and 1980, the relative immigrant wage declined by about 12 percentage points. Similarly, recent immigrants in 1960 earned 12.6 percent less than natives aged 25–44. By 1970, the wage disadvantage had increased to 14.8 percent, and by 1980 to 25.7 percent. Between 1960 and 1980, the immigrant relative wage had fallen by 13 percentage points. In Panel A of Table 1, the decline in the relative immigrant wage over the 1960–80 period was 17 percent. It seems that accounting for differential period effects between immigrants and natives only slightly attenuates the downward trend in immigrant skills and labor market performance.

II. The Role of National Origin

Because of changes in immigration policy and in economic and political conditions both in the United States and abroad, the national origin mix of the immigrant flow during the 1970's was different than that of earlier waves. In the 1950's, over half of the immigrants originated in Europe, 39 percent originated in the Western Hemisphere, and 6 percent originated in Asia. In the 1970's, 18 percent originated in Europe, 44 percent originated in the Western Hemisphere, and 35 percent originated in Asia. There is substantial dispersion in the skills and labor market performance of different national origin groups (see my 1987 article). The results presented below indicate that the changing national origin mix of the immigrant flow largely explains the decline in skills documented in the previous section.

The immigrant population is divided into 41 national origin groups, as well as a residual "other" category. (My 1987 article lists the 41 countries.) The 41 groups account for over 90 percent of the 1951–80 immigrant flow. For each of the 42 groups (including other immigrants), I calculated the average educational attainment and log wages.

Let Y_t be the average value for a particular characteristic observed in the recent im-

migrant population in year t (relative to that observed among natives aged 25–64). By definition, Y_t can be written as

$$(1) \quad Y_t = \sum_j p_{jt} y_{jt},$$

where y_{jt} is the average value observed among recent immigrants from national origin group j in year t (relative to natives); and p_{jt} is the fraction of the immigrant flow in year t originating in country j .

It is useful to define the value that would have been observed if a different national origin mix had migrated to the United States, such as the national origin mix observed at time τ , $p_{j\tau}$:

$$(2) \quad Y(t, \tau) = \sum_j p_{j\tau} y_{jt}.$$

The impact of a changing national origin mix is then given by

$$(3) \quad Y_t - Y(t, \tau) = \sum_j y_{jt} (p_{jt} - p_{j\tau}).$$

I wish to determine the extent to which the changing national origin mix explains the decline in immigrant skills. In other words, how important is the change predicted by equation (3) in terms of the total change? Because the data on y_{jt} can be chosen arbitrarily from any of the censuses, there are a number of answers to this question. To easily summarize the evidence, I use the vector y_{jt} estimated from the 1980 Census. The selection of alternative vectors does not alter the qualitative nature of the results.

Table 2 reports the results of simulations using this framework.⁵ Consider the evidence regarding educational attainment. As I documented earlier, the average schooling of recent immigrants (relative to natives) fell by 1.1 years between 1960 and 1980, and by .4 years between 1970 and 1980. Using

⁵Because of the small number of immigrants during the 1930's, Table 2 focuses on explaining the changes observed between 1960 and 1980.

TABLE 2—DECOMPOSITION OF CHANGES IN IMMIGRANT SKILLS (RELATIVE TO NATIVES)

Variable	Change: 1960–1980		Change: 1970–1980	
	(1)	(2)	(1)	(2)
<i>Education</i>	–1.075	–.932	–.442	–.215
<i>Log Wage</i>	–.171	–.213	–.139	–.066
<i>Adj. Log Wage</i>	–.111	–.147	–.075	–.024

Note: Col. (1) is total change; Col. (2) is change due to national origin.

equation (3), I estimate that the observed changes in national origin would have led to a decline in the educational attainment of $-.9$ years between 1960 and 1980, and $-.2$ years between 1970 and 1980. Put differently, at least half of the drop in educational attainment can be accounted for by changes in the national origin mix of immigrant flows.

The data also indicate that the relative wage of newly arrived immigrants fell by 17.1 percentage points between 1960 and 1980, and by 13.9 percentage points between 1970 and 1980. I estimate that changes in national origin alone would have caused a drop in the relative immigrant wage of 21.3 percentage points between 1960 and 1980, and of 6.6 percent between 1970 and 1980. Changes in national origin, therefore, explain at least two-thirds of the wage decline over this period.

Finally, the data show an 11.1 percentage point drop in the adjusted immigrant wage between 1960 and 1980, and a 7.5 percentage point drop between 1970 and 1980. The changing national origin mix of immigrants account for over a third of the decline in adjusted wages.

In sum, shifts in the national origin mix of the immigrant flow are responsible for a substantial decline in immigrant skills and for a deterioration in the labor market performance of successive immigrant waves over the postwar period. In effect, the typical immigrant originating in the source countries responsible for immigration in the

1970's is less skilled than the typical immigrant originating in the source countries responsible for immigration in the 1950's and early 1960's.

III. Skill Transmission Across Immigrant Cohorts

Consider the regression model:

$$(4) \quad y_{jt} = \alpha_0 + \alpha_1 y_{j,t-1} + \varepsilon_{jt},$$

where y_{jt} is the educational attainment or wage of an immigrant cohort from country j that arrived in the United States at time t . The specification in equation (4) raises the question of whether immigrant skills are transmitted across cohorts. This transmission may occur because most immigrants arriving in the United States already have relatives residing here. It is likely that information regarding the U.S. labor market is relayed across the links in the immigration chain. This would suggest that immigrant cohorts that were particularly successful in the U.S. labor market would be followed by cohorts that would also be relatively successful.

The problem with testing the hypothesis underlying equation (4) is that a positive estimate of α_1 may have nothing to do with the transmission of skills, and may be entirely due to a fixed effect that characterizes the national origin group. The correct statistical model is then given by

$$(5) \quad y_{jt} = \alpha_0 + \alpha_1 y_{j,t-1} + v_j + \delta_{jt},$$

where v_j is the national-origin fixed effect.

For each of the 41 countries in the analysis, I have information on the skills and wages (relative to natives) of three successive waves: the 1955–59, the 1965–69, and the 1975–79 arrivals. The data is pooled across censuses, so that the skills of the 1965–69 immigrants are related to those of the 1955–59 wave, while the skills of the 1975–79 immigrants are related to those of the 1965–69 wave. The pooling yields 82 observations and allows the estimation of a fixed-effects model. Because of period ef-

TABLE 3—SKILL TRANSMISSION ACROSS IMMIGRANT COHORTS

Dependent Variable	y_{t-1}	Controls for Fixed Effects	R^2
Education	.8690 (16.54)	No	.780
Education	.2081 (1.29)	Yes	.933
Log Wage	.8230 (11.29)	No	.633
Log Wage	.2911 (2.46)	Yes	.887
Adj. Log Wage	.6105 (7.59)	No	.466
Adj. Log Wage	.1454 (1.65)	Yes	.885

Note: The t -ratios are shown in parentheses. All regressions also include an intercept and a dummy variable indicating if the dependent variable was drawn from the 1980 Census.

fects, the regression also includes a dummy variable indicating if the dependent variable is drawn from the 1970 or 1980 Census.

Table 3 presents the estimated regressions. Without controlling for the fixed effects, there exists a strong positive correlation in Education, Log Wages, and Adjusted Log Wages across cohorts. After controlling for fixed effects, there remains a positive correlation between Wages and Adjusted Log Wages across cohorts, and an insignificant relationship between educational attainment across cohorts. This pattern of coefficients suggests that there is skill transmission across cohorts. After all, educational attainment at the time of ar-

rival (for prime-age men) is difficult to transmit, and the fixed effect should account for the correlation in schooling across successive waves. In contrast, labor market success in the United States (as measured by the wage rate) can be transmitted, and the evidence indicates that it is. Note also that the transmission coefficient is less than unity, indicating a regression towards the mean across successive waves.

In view of the relatively small number of observations and of the importance of fixed effects, it is best to interpret the results with caution. They do suggest, however, that the future study of the transmission of skills within immigrant households may provide important insights into the labor market performance of the foreign-born in the United States.

REFERENCES

- Borjas, George J., "Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants," *Journal of Labor Economics*, October 1985, 3, 463-89.
- , "Self-Selection and the Earnings of Immigrants," *American Economic Review*, September 1987, 77, 531-53.
- , *Friends or Strangers: The Impact of Immigrants on the U.S. Economy*, New York: Basic Books, 1990.
- Greenwood, Michael J. and McDowell, John M., "The Factor Market Consequences of Immigration," *Journal of Economic Literature*, December 1986, 24, 1738-72.

Immigration and Wages: Evidence from the 1980's

By KRISTIN F. BUTCHER AND DAVID CARD*

More immigrants entered the United States during the past decade than in any comparable period since the 1920's. Among the issues raised by this influx, none is as controversial as its effect on the labor market opportunities of native-born workers. Evidence on the labor market consequences of immigration is limited (see Michael Greenwood and John McDowell, 1986, and George Borjas, 1990). This paper presents new evidence on the effects of immigration, based on changes in the distributions of wages in 24 major cities during the 1980's. Although immigrant inflows are small relative to the populations of most cities, recent immigrants are a significant fraction of less-educated workers in many cities. We therefore concentrate on measuring the effects of immigration at the lower tail of the wage distribution. In particular, we ask whether recent declines in the real earnings of the least-skilled workers in the U.S. economy are related to immigration. Our empirical analysis reveals large differences across cities in the relative growth rates of wages for low- and high-paid workers. Nevertheless, these differences bear little or no relation to the size of immigrant inflows. Our results therefore confirm the findings of earlier studies, based on 1970 and 1980 Census data, that suggest that the labor market consequences of higher immigration are relatively small.

I. Characteristics of New Immigrants in 1980 and 1985

A standard approach to measuring the labor market effects of immigration is to treat different cities within the United States as distinct labor markets, and to compare labor market outcomes across cities with higher and lower immigrant densities. We

follow this approach here, tracking wages in 24 major cities during the period from 1979 to 1989. Our sample includes the 10 most immigrant-intensive cities identifiable in *Current Population Survey (CPS)* microdata files, along with a group of 14 other cities (listed in Table 2). The other cities were selected by a variety of criteria, including stable boundaries, relatively large sample sizes, and a desire for geographic comparability with the high-immigrant cities. We believe that the sample gives a fair picture of the variation in immigrant inflow rates and relative immigrant "quality" across major U.S. cities.

We begin by presenting some simple evidence on the nature of recent immigrant inflows into the United States. The first two columns of Table 1 contain data from the 1980 Census on the characteristics of natives and recent immigrants (i.e., those who immigrated between 1975 and 1980) in the 24 cities in our sample. Recent immigrants are younger, less educated, and more likely to be male than natives (or earlier immigrants). The proportion of Hispanics is also much higher among recent immigrants than in the native population. The education distribution of newly arriving immigrants is relatively disperse: the fraction of college graduates is about the same as in the native population, but close to one-quarter of recent immigrants have less than an elementary education. As a result, recent immigrants make up 17 percent of the population in these cities with 6 or fewer years of schooling, and 10 percent of the population with less than an eighth-grade education.

Columns 3 and 4 of Table 1 present the characteristics of individuals in the March 1985 *CPS*, classified by whether or not the respondent was living in the United States 5 years earlier. Based on responses to a similar question in the 1980 Census, we estimate that 85 percent of those living abroad 5 years ago are immigrants. Demographic differences between the two groups

*Princeton University, Princeton NJ 08544.

TABLE 1—COMPARISON OF RECENT IMMIGRANTS
AND OTHERS IN 24 CITIES,
1980 CENSUS AND MARCH 1985 CPS

	1980 Census		March 1985 CPS	
	(1)	(2)	(3)	(4)
Pct of Total	86.8	3.4	96.6	3.4
Pct Female	51.4	48.2	51.5	46.5
Mean Age	37.4	31.3	37.7	31.7
Pct Age 16–24	24.7	34.4	21.7	33.0
Pct Hispanic	5.0	39.0	11.2	36.7
Education (Years):				
Mean	12.5	10.7	12.7	11.3
Pct 0–6	2.3	22.8	3.4	15.0
Pct 16+	18.0	17.3	22.3	21.0
Work in Previous Year:				
Pct Worked	71.5	59.5	71.4	60.6
Avg. Log Wage	1.82	1.51	2.06	1.75
Std. Deviation	0.66	0.66	0.66	0.64

Note: Samples contain individuals age 16–68 in 24 cities. Col. (1) contains natives; Col. (2) contains immigrants who entered the United States between 1975 and 1980; Col. (3) contains individuals who were living in the United States in March 1980; Col. (4) contains individuals who were living abroad in March 1980.

are consistent with the differences between recent immigrants and others in the 1980 Census. The 30 percent wage gap between the new arrivals and other workers is also very similar to the gap between recent immigrants and others in 1980. We conclude from these comparisons that the relative “quality” of arriving immigrant cohorts was relatively stable between 1975 and 1985.

II. Immigration to Specific Cities

From this general overview we turn to a specific examination of the nature of immigrant flows to each of the 24 cities in our data set. Table 2 provides information on the percent of “recent immigrants” in each city in 1980 and 1985, together with data on the overall population growth rate between 1980 and 1987, and the wage gap between recent immigrants and natives in 1980. Two important features of U.S. immigration are highlighted in the table. First, recent immigrants are highly concentrated in only a few cities. Three cities (New York, Los Angeles, and Miami) accounted for 51 percent of recent immigrants in both 1980 and 1985. Second, there is substantial variation across

TABLE 2—CHARACTERISTICS OF 24 MAJOR CITIES

	Percent Recent Imms:		Growth Rate ^a	Wage Gap ^b
	1980	1985		
New York	4.4	6.0	0.4	34.5
Los Angeles	7.7	7.3	1.8	44.5
Chicago	2.5	2.4	0.3	32.3
Philadelphia	0.9	1.5	0.5	25.9
Detroit	0.8	0.9	–0.4	15.2
San Francisco	4.3	4.1	1.3	27.5
Washington, D.C.	2.8	3.8	1.6	27.7
Baltimore	0.7	1.6	0.7	26.6
Houston	3.3	3.5	2.4	32.2
Minneapolis	1.1	0.9	1.3	3.4
Dallas	1.7	4.7	3.4	31.6
Seattle	1.9	2.5	1.6	24.3
Anaheim	4.8	3.9	2.0	44.2
Milwaukee	0.6	3.3	–0.1	19.4
Atlanta	0.7	2.4	3.1	17.1
San Diego	4.0	3.9	2.9	34.8
Miami	6.1	8.7	1.4	30.8
Denver	1.4	0.9	2.0	9.2
Riverside, CA	1.9	4.4	4.4	22.2
San Jose	4.4	3.6	1.3	20.7
New Orleans	1.3	2.1	0.7	14.5
Tampa	0.7	5.1	2.8	21.7
Portland	1.5	1.6	0.8	14.2
Sacramento	1.7	3.2	2.8	17.0

^aAnnual percentage growth rate in population 1980–87.

^b1980 wage gap between recent immigrants and native born.

cities in the composition of recent immigrant inflows. As a general rule, the quality of recent immigrants, measured by their wage gap relative to native workers, is lower in cities with higher inflow rate. A key correlate of both inflow rates and the relative wage of recent immigrants is the fraction of Hispanic immigrants, which ranges from under 10 percent in Detroit, Minneapolis, Seattle, and Portland to over one-half in cities in California and Texas.

An important feature of immigration to a local labor market is its effect on population growth. A natural assumption is that an inflow of new immigrants generates a proportional increase in the labor force and population of a city. Recent research by Randall Filer (1990), however, suggests that the intercity migration decisions of native workers are highly sensitive to immigrant inflows. Indeed, Filer’s analysis of population movements between 1975 and 1980 im-

plies that immigrant arrivals are almost completely offset by native outflows.

There is some evidence of offsetting out-migration in Table 2, particularly for the high-immigrant cities of New York, Los Angeles, and Miami. All three cities had large immigrant inflows but relatively modest growth rate during the 1980's. In the absence of out-migration, an increase in the fraction of new immigrants will raise the population growth rate of a city point-for-point. Therefore, if native inflow rates are independent of immigration rates, population growth rates should be linearly related to immigration inflow rates, with a slope of 1.0. In fact, the slope of a regression line fitted to all 24 cities in our data set is 1.04 (with a standard error of 0.54). When a similar regression is fit to the subset of observations that excludes New York, Los Angeles, and Miami, however, the estimated slope is much higher (2.76, with a standard error of 0.67). This regression accounts for about one-half of the variation in growth rates in the subset of 21 cities. From this evidence we conclude that native in-migration flows during the 1980's were actually *positively* correlated with inflows of recent immigrants to all but the 3 most immigrant-intensive cities.

One explanation for the difference between the highest-immigrant cities and other major cities is based on the composition of immigrant inflows. Between 1980 and 1985, over one-half million Cuban and Southeast Asian refugees arrived in the United States (Frederick Hollmann, 1990, Table V). Most of the Cubans, and perhaps one-half of the Asians, settled in either New York, Los Angeles, or Miami. To the extent that the refugees were drawn to these cities by cultural and ethnic ties, their location decisions may have been less sensitive to local labor market conditions than the decisions of other newly arriving immigrants. It should be pointed out, however, that New York and Los Angeles had large immigrant inflows during the 1970's and grew more slowly than most other U.S. cities between 1970 and 1980. The experiences of these two cities during the 1980's were therefore in keeping with earlier trends. What is appar-

ently different between the 1970's and 1980's is the emergence of a positive relation between immigration and overall population growth among other cities.

III. Immigration and Wages

Table 3 turns to an examination of wage outcomes in different cities during the past decade. The data are taken from merged files of the 12 monthly *Current Population Surveys* administered in 1979, 1980, 1988, and 1989, and pertain to hourly wage rates (for hourly rated workers) or the ratio of average weekly earnings to average weekly hours (for salaried workers). For each city and each year, we have calculated the 10th and 90th percentiles of the log wage distribution. The first two columns of Table 3 represent (unweighted) averages of the 1979 and 1980 percentiles, and the second two columns represent changes from the 1979-80 average to the 1988-89 average.

TABLE 3—LEVELS AND CHANGES IN PERCENTILES OF LOG WAGES IN 24 CITIES: 1979-80 TO 1988-89

	Percentiles in 1979-80		Changes from 1979-80 to 1988-89	
	10th	90th	10th	90th
New York	1.14	2.39	0.47	0.67
Los Angeles	1.14	2.46	0.34	0.56
Chicago	1.18	2.50	0.30	0.48
Philadelphia	1.12	2.38	0.44	0.59
Detroit	1.13	2.51	0.29	0.46
San Francisco	1.26	2.56	0.35	0.56
Washington, D.C.	1.18	2.69	0.43	0.41
Baltimore	1.12	2.39	0.36	0.54
Houston	1.14	2.45	0.22	0.52
Minneapolis	1.15	2.47	0.39	0.50
Dallas	1.14	2.35	0.28	0.58
Seattle	1.27	2.50	0.29	0.49
Anaheim	1.14	2.55	0.40	0.50
Milwaukee	1.12	2.39	0.30	0.46
Atlanta	1.12	2.40	0.39	0.58
San Diego	1.12	2.44	0.36	0.52
Miami	1.10	2.20	0.29	0.63
Denver	1.15	2.50	0.29	0.45
Riverside, CA	1.12	2.41	0.37	0.58
San Jose	1.25	2.63	0.40	0.58
New Orleans	1.13	2.36	0.19	0.49
Tampa	1.10	2.17	0.29	0.59
Portland	1.19	2.46	0.29	0.44
Sacramento	1.12	2.46	0.43	0.48

The distribution of hourly wages in each city is approximately lognormal, although the distributions contain prominent "spikes" at points like \$3.00, \$5.00, and \$10.00 per hour. The spread between the 10th and the 90th percentile of log wages in any city tends to be strictly proportional to the estimated standard deviation, as is the case for a normal distribution.

The 1979–80 data reveal sharp differences across cities in both the level and dispersion in wages. Interestingly, there is less variation across cities in the 10th percentile of wages than in wages at the middle or upper end of the earnings distribution. This is apparently due to the restraining effect of the minimum wage, that served as the 10th percentile of wages in many cities in 1979 and 1980. One implication of a binding national wage floor is that the dispersion of wages within a city is highly correlated with the average level of wages. In 1979–80, this pattern is clearly present in the data. Despite the differences in mean log wages across cities, however, 97 percent of the overall variation in individual wages for workers in our sample of cities is within-city variation.

Intercity differences in the 10th percentile of wages in 1979–80 are uncorrelated with differences in the fraction of recent immigrants (or total immigrants) in the city. Wages at the upper end of the earnings distribution are weakly negatively related to the fraction of recent immigrants, but weakly positively related to the overall fraction of immigrants. These small and unsystematic correlations are consistent with findings in the previous literature. Looking across cities in 1979–80, there is no evidence of any effect of immigration on the level of wages.

Changes in the distribution of wages over the last decade also show considerable intercity variation. Mean log wages in most cities grew at roughly the same rate as the Consumer Price Index (CPI), which rose 44 percent between 1979–80 and 1988–89. Mean log wages in Detroit, Houston, and New Orleans, however, grew much more slowly than average consumer prices, while those in New York grew faster. As shown in the third and fourth columns of Table 3,

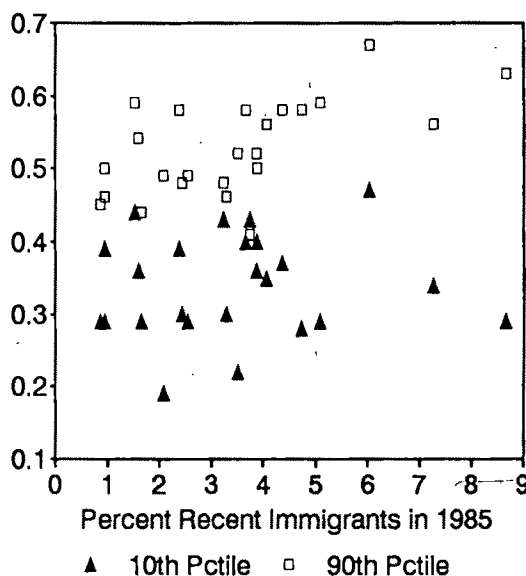


FIGURE 1. IMMIGRATION AND CHANGE IN WAGE DECILES IN 24 MAJOR CITIES, 1979–80 TO 1988–89

there are significant differences across cities in the relative growth of the 10th and 90th percentiles of wages (and in the corresponding change in the standard deviation of wages). In every city except Washington, D.C., wage rates in the upper tail of the earnings distribution grew more quickly than those in the lower tail. Thus, the growth in wage inequality during the 1980's (Chinhui Juhn et al., 1989) was almost entirely due to an increase in the within-city variance.

The rise in wage inequality over the 1980's was greater in cities with relatively bigger immigrant inflows. This is illustrated in Figure 1, which plots the changes in the 10th and 90th percentiles of wages for each city against the fraction of recent immigrants in 1985. Contrary to our expectations, however, the data suggest that higher immigration is associated with more rapid increases in the 90th percentile of wages, rather than with any relative decline in the 10th percentile of wages.

One explanation for the pattern of the data in Figure 1 is that the cost of living rose more rapidly in cities with larger immigrant inflows, and that these increases in

prices lead to wage increases for high-wage workers. To test this possibility we constructed wage changes relative to city-specific cost-of-living indexes, which are available from the Bureau of Labor Statistics for most of the cities in our sample. As hypothesized, the change in the city-specific cost of living is positively correlated with the fraction of recent immigrants in 1985. Adjusting for changes in the local cost of living, the change in the 90th percentile of wages is still positively related to the fraction of recent immigrants, but the regression coefficient is halved and falls to statistical insignificance. Similarly, an adjustment for the local cost of living causes the regression coefficient relating the change in the 10th percentile of wages to the fraction of recent immigrants in 1985 to become (slightly) negative. Thus, adjusting for city-specific changes in the cost of living, we find less evidence of a positive correlation between immigration rates and the growth of high-skilled wages, and more evidence of a negative correlation between immigration and the growth of low-skilled wages. Neither of these correlations, however, is large or statistically significant.

We have also calculated the effect of higher immigration on the various percentiles of the wage distribution, controlling for such factors as the overall population growth rate, the fraction of immigrants initially living in each city, and the initial level of wages in the city. In no case do we find a large or statistically significant effect of immigration on the rate of increase of wages for the least-skilled workers.

IV. Conclusions

We believe that the evidence we have assembled for the 1980's confirms the conclusions from earlier studies of 1970 and

1980 Census data. In particular, we find little indication of an adverse wage effect of immigration, either cross sectionally or within cities over time. Even for workers at the 10th percentile of the wage distribution there is no evidence of a significant decline in wages in response to immigrant inflows. The one important difference that emerges between our analysis and earlier studies is the finding of a positive link between immigrant inflows and net native migration. During the 1980's, rapidly growing cities in Texas, Florida, and California attracted both native and newly arriving immigrant workers. This configuration is quite different from the pattern of offsetting immigrant and native population inflows identified by Filer in earlier data.

REFERENCES

- Borjas, George, *Friends or Strangers: The Impact of Immigrants on the U.S. Economy*, New York: Basic Books, 1990.
- Filer, Randall K., "The Impact of Immigrant Arrivals in Migratory Patterns of Native Workers," in R. Freeman and G. Borjas, eds., *U.S. Immigration: Destinations and Sources*, Chicago: University of Chicago Press, 1990.
- Greenwood, Michael and McDowell, John, "The Factor Market Consequences of U.S. Immigration," *Journal of Economic Literature*, December 1986, 24, 1738-72.
- Hollmann, Frederick W., "United States Population Estimates, by Age, Sex, Race, and Hispanic Origin: 1980 to 1988," *Current Population Reports*, Series P-25, No. 1045, Washington: USGPO, 1990.
- Juhn, Chinhui, Murphy, Kevin M., and Pierce, Brooks W., "Wage Inequality and the Rise in Returns to Skill," unpublished manuscript, University of Chicago Graduate School of Business, 1989.

Immigrants in the American Labor Market: Quality, Assimilation, and Distributional Effects

By ROBERT J. LALONDE AND ROBERT H. TOPEL*

Immigration to the United States during the 1970's and 1980's was greater than in any decade since the 1920's. Simultaneously, the proportion of immigrants arriving from Europe or English-speaking countries declined dramatically from 46 percent in the 1960's to only 13 percent in the 1980's. These changes in the pace and composition of immigrant flows have raised three main concerns that have dominated recent literature and policy discussions. First, are recent immigrants less prepared to succeed in the U.S. labor market than their predecessors? The typical new immigrant may bring skills (including language, culture, and educational attainment) that are less attuned to the American market. Further, this decline in immigrant "quality" may have been magnified by immigration reforms of the 1960's, that gave less weight to individuals' skills in admission decisions. Thus average immigrant quality may have also declined *within* ethnic groups. Second, do new immigrants recover from their initial earnings disadvantage? As a policy issue, if new immigrants do not assimilate, the increased immigrant flows may place additional burdens on public welfare systems, while exacerbating other social problems associated with persistent poverty.

The third concern is the distributional impact of immigration on natives' welfare. To the extent that immigrant and native skills are substitutable, increased immigration may reduce natives' earnings and employment prospects. In fact, the 1970's and 1980's were periods of declining real wages and rising unemployment among less-skilled

American workers. These facts have motivated policy proposals and legislation designed to curtail the entry of low-wage immigrants to the United States, while giving greater emphasis to skill-based criteria in admissions decisions.

This paper provides evidence on immigrants' performance and impact in the U.S. labor market. We document that new immigrants do bring fewer marketable skills to the United States than did earlier cohorts, and that changes in the source countries of recent immigrants account for all of this decline in immigrant "quality." We find no important evidence that quality has declined *within* immigrant ethnic groups. We also show that immigrants assimilate rapidly in the U.S. market (10 years of U.S. experience offsets most of the earnings disadvantage of new immigrants), and that assimilation is more rapid for groups who start with lower initial wages.

These findings imply that immigrants' long-run earnings potential is similar to that of ethnically similar natives. In this sense, we think that fears about declining immigrant quality have been exaggerated. We present additional evidence that increased immigration has had a negligible impact on U.S. workers' wages and employment prospects. Taken together, these results suggest that recent changes in the rate and composition of immigration will not have serious long-run effects on the U.S. labor market.

I. Changes in Immigrant Quality

Recent studies (see George Borjas, 1985; Barry Chiswick, 1986) indicate a decline in the labor market skills of recent immigrants to the United States. There are two possible sources for this decline. First, non-European immigrants may arrive with fewer skills. Second, even absent changes in source

*University of Chicago, Chicago, IL 60637, and NBER. We gratefully acknowledge the support of the Alfred P. Sloan Foundation and the National Science Foundation. Topel's work was supported by the William Ladany Faculty Research Fund at the Graduate School of Business, University of Chicago.

TABLE 1—DIFFERENCES IN WAGES AND SCHOOLING FOR RECENT IMMIGRANTS AND NATIVES

Immigrant Group Years in U.S.	Wage Relative to Natives		
	1970 (1)	1980 (2) ^b	1980 (3)
A: Immigrants' Relative Log Weekly Wages^a			
All:			
1-5	-.20	-.35	-.35
6-10	-.06	-.25	-.21
11-15	0	-.14	-.14
European:			
1-5	0	-	-.02
6-10	.07	-	-.06
11-15	.07	-	0
Asian:			
1-5	-.09	-	-.22
6-10	.09	-	-.04
11-15	.11	-	.15
Mexican:			
1-5	-.68	-	-.69
6-10	-.49	-	-.59
11-15	-.34	-	-.48
B: Immigrants' Years of Completed Schooling^c			
	1970	1980	
All Immigrants:	10.8	11.6	
Europeans:	10.9	12.1	
Asians	14.4	14.3	
Mexicans	6.4	7.0	
Natives	11.6	12.7	

Source: Calculations from 1970 and 1980 Census.

^aAverage differences between log weekly wages of immigrants and natives with comparable labor market experience.

^bThe wage differentials in this column are weighted by the 1980 distributions of immigrant shares and native education.

^cAverage years of completed schooling among immigrants with 0-10 years in the United States and native male labor force participants.

and 1980, new immigrants' schooling increased overall, but declined relative to natives. Simultaneously, new immigrants' earnings disadvantage rose by an additional 15 points, to 35 percent (col. 3). In this sense, the average (relative) "quality" of new immigrants declined during the decade, even though average schooling levels suggest no decline in quality *within* immigrant groups.

The evidence in Table 1 suggests that change in source countries and changes in average years of native schooling may account for the decline in relative immigrant earnings. To test that proposition, we recalculated the relative earnings of 1970 immigrants, but we weighted each immigrant group by the group's relative share among new immigrants in 1980. We also reweighted natives' earnings to reflect their 1980 distribution of completed schooling. The calculations show that if the distribution of immigrants across source countries had been the same in 1970 as in 1980, and if educational attainment of U.S. workers had been the same in 1970 as in 1980, the relative earnings of immigrants in 1970 would have been at their 1980 level. For example, persons in the United States less than 5 years in 1970 would have earned 35 percent less than comparable natives, which is identical to the 1980 earnings gap. This indicates that the main factor affecting the relative earnings of new immigrants is simply where they come from. There is no evidence of declining quality *within* immigrant ethnic groups.

II. Assimilation of Immigrants

countries, average productivity may decline *within* immigrant groups when changes in laws and incentives favor the entry of less-skilled individuals.

To address the question of declining immigrant quality, Table 1 compares recent immigrants' and natives' earnings and education in the 1970 and 1980 Censuses. In 1970, the average new immigrant arrived in the United States with 10.8 years of completed schooling, and earned about 20 percent less than natives with similar labor market experience (col. 1). Between 1970

The evidence in Table 1 suggests that average immigrant quality (as measured by earning capacity) declined in the 1970's. Since Asia and Latin America remained the most important sources of (legal and illegal) immigration during the 1980's, there is little doubt that the decline in the earning capacity of new immigrants has continued. But since immigrants assimilate with time in the United States, the decline in new immigrants' *initial* earning capacity overstates the long-term decline in immigrant quality. Further, those groups with the largest initial

earnings disadvantage assimilate the most. Thus an increase in the shares of Asians and Hispanics among new immigrants reduces new immigrants' relative earnings, but increases the average rate of assimilation.

These points are also demonstrated in Table 1. The cohort of immigrants who arrived in the United States between 1965 and 1969 (1–5 years in the United States in 1970) earned 20 percent less than comparable natives in 1969. By 1979, that cohort had been in the United States for 11–15 years and they earned 14 percent less than comparable natives, so relative earnings grew by 7 percent over the decade. This growth is the average of assimilation rates within immigrant groups. The table shows that assimilation is much larger among Asians and Hispanics than among Europeans, who were the largest immigrant group in 1970. Europeans who arrived between 1965 and 1969 experienced no relative earnings growth at all during the 1970's, but they also started from parity in 1969. By comparison, Asians and Mexicans experienced relative earnings growth of 24 and 20 percent over the decade, respectively. Since Asians and Mexicans accounted for vastly larger proportions of new immigrants in the 1970's and 1980's, the rate of convergence between immigrant and native earnings will be correspondingly larger than in the past.

Rapid earnings growth does not mean that assimilation eliminates the earnings gap between immigrants and natives. Nor should we expect the gap to be eliminated, since there are important earnings differentials among *natives* of different ethnic backgrounds. To illustrate, column 1 of Table 2 reports estimated wage differentials among natives of various ancestry, based on cross-sectional data from the 1980 Census. Since these estimates control for the usual list of observable background characteristics, the reported values are approximate percentage differentials between the earnings of the indicated ethnic group and European-Americans with the same observable characteristics. As the table shows, a typical Mexican-American earns about 11 percent less than a comparable European, while an Asian-American earns about 7 percent less.

TABLE 2—RATES OF IMMIGRANT ASSIMILATION

Ethnic Group	(1)	(2)	(3)	(4)	(5)
Europeans	–	–.05 (.02)	.09 (.02)	.08 (.03)	.08 (.03)
Asians	–.07 (.03)	–.33 (.02)	.24 (.03)	.25 (.03)	.24 (.09)
Middle Easterners	.06 (.05)	–.40 (.06)	.28 (.07)	.29 (.07)	.42 (.20)
Mexicans	–.11 (.02)	–.22 (.02)	.17 (.03)	.22 (.03)	.21 (.09)
Other Latin American and Caribbean	–.20 (.03)	–.22 (.03)	.23 (.03)	.24 (.03)	.19 (.09)

Source: Calculations using 1970 and 1980 U.S. Census Microdata files and our paper (1991b, Table 5).

Note: Col. (1) = log wage differences for ethnic natives; Col. (2) = initial gap with ethnic natives; the rates of assimilation in cols. (3), (4), and (5) are relative to: Col. (3) ethnic natives 1980 cross section; Col. (4) old immigrants 1980 Cross Section; Col. (5) old immigrants 1970–80 panel.

Estimates control for years of completed schooling, a quartic in experience, and interactions between schooling and experience. Additional controls do not affect the results. Assimilation rates estimate the effect of the first 10 years' U.S. experience on wages, measured relative to observationally identical ethnic natives and ethnic immigrants who have been in the U.S. for more than 30 years. Standard errors are in parentheses.

In our view, these estimates suggest that the relevant question is not whether immigrants catch up with the modal native (who has European ancestry), but rather whether they catch up to natives of similar ancestry to their own.

To frame this question more precisely, consider the following decomposition of the regression-adjusted wage differential between immigrants from arrival cohort i in census year t and natives of similar ethnicity:

$$(1) \quad \varepsilon_{it} = a_{it} + b_{it} + u_i,$$

where the parameters a_{it} represent the average level of accumulated, U.S.-specific, human capital embodied in members of arrival cohort i during year $t = 1970$ or $t = 1980$. Assimilation occurs when the relative human capital of an immigrant cohort rises with time spent in the United States, $a_{it} < a_{i,t+10}$. Identification of this from estimates

of ε_{it} is complicated by the appearance of b_{it} , that represents the impact of aggregate labor market conditions on immigrant cohort i 's earnings, and by u_i , that represents the cohort-average value of other unobserved factors (immigrant "quality") that affect productivity. Based on (1), two possible estimators of immigrant assimilation are possible. From a single cross section, say $t = 1980$, we can estimate the effect of 10 years residence in the United States by comparing the relative earnings of persons who arrived between 1975 and 1979 ($i = 75$) to the earnings of similar immigrants who arrived between 1965 and 1969 ($i = 65$):

$$(2) \quad \varepsilon_{65,80} - \varepsilon_{75,80} = a_{65,80} - a_{75,80} + b_{65,80} - b_{75,80} + u_{65} - u_{75}.$$

This provides unbiased estimates of assimilation only if (i) there are no time effects on relative earnings of the two cohorts, $E(b_{65,t} - b_{75,t}) = 0$, and (ii) the cohorts have the same average talent, $E(u_{65} - u_{75}) = 0$. For example, when the average quality of successive immigrant cohorts declines, $u_{65} > u_{75}$, equation (2) will overstate the rate of immigrant assimilation. This issue underlies Borjas's 1985 criticism of Chiswick's 1978 estimates of assimilation.

The alternative to (2) is to form a quasi panel by following the wage growth of a single cohort between the 1970 and 1980 Censuses. For the cohort that arrived between 1965 and 1969, this estimate is

$$(3) \quad \varepsilon_{65,80} - \varepsilon_{65,70} = a_{65,80} - a_{65,70} + b_{65,80} - b_{65,70}.$$

This estimator of assimilation will be unbiased if aggregate market conditions do not change the cohort's relative earning power, $E(b_{65,80} - b_{65,70}) = 0$. Our estimates of (3) use earlier immigrants and natives of the same ethnicity as recent immigrants as normalizing groups. If recent immigrants' human capital substitutes for that of ethnically similar earlier immigrants or natives, the identifying condition is likely to be satisfied.

As shown by Table 2, both estimates of assimilation indicate that most of new immigrants' earnings disadvantage is eliminated after only 10 years' experience in the U.S. labor market. Column 2 shows the initial difference in log weekly wages between immigrants who arrived between 1975 and 1979 and natives of the same ethnicity. New immigrants' wages typically start out well below those of natives. But the data show that this earnings disadvantage rapidly narrows. Column 3 and 4 show estimates of 10 years' assimilation, based on applying (2) to cross-sectional data from the 1980 Census. Although the two sets of estimates rely on different normalizing groups, they nevertheless indicate rapid and nearly identical rates of immigrant assimilation. Both estimates indicate that Asian immigrants will overcome 75 percent ($= .24/.32$) of their original wage shortfall after 10 years in the United States. Other groups show similar patterns of strong assimilation. Notice that the largest wage gains occur for groups with the largest initial wage gap.

These estimates are subject to the criticism that immigrant quality may have fallen over time within ethnic groups. That decline would cause (2) to overstate the rate of assimilation. Although our estimates in Table 1 showed no evidence of declining immigrant quality within ethnic groups, the "panel" estimator in (3) avoids this concern by following the earning growth of a fixed cohort over time. As shown in column 5, panel estimates confirm that declining immigrant quality within ethnic groups is not an important issue: estimates of assimilation are large, and they are nearly identical to the cross-sectional estimates shown in column 4. We conclude that immigrants assimilate rapidly in the U.S. labor market.

III. The Effects of Increased Immigration on Wages and Employment

Immigrant flows accounted for only about 5 percent of growth in aggregate labor supply during the 1970's and 1980's. At this level of aggregation it is difficult to argue that immigration had important effects on natives' wages and employment prospects.

Yet immigration is heavily concentrated in certain geographic areas and among certain skill groups. For example, during the 1970's immigrants accounted for nearly two-thirds of labor force growth in Los Angeles, and over one-third of growth in Miami. This concentration raises the possibility of significant distributional effects on some natives' wages, at least in the short run. In light of these facts, we have argued elsewhere (1991a) that the largest impact of immigration on wages and employment must be for individuals who are good substitutes for new immigrants in terms of both location and skills. In our analysis, these individuals are other current and past immigrants of similar ethnicity, as well as less-skilled natives located in areas that have experienced large immigration flows.

As shown by Table 3, our estimates indicate that increased immigration has had relatively small effects on recent immigrants', and on young black and Hispanic natives', wages and earnings. The econometric framework for isolating these effects is spelled out in our 1991a paper. Briefly, the estimates are derived by comparing the *relative* wages of the indicated groups among SMSAs with varying amounts of immigration, while controlling for a host of background characteristics. As above, there are two approaches to estimating these effects. The first column of Table 3 shows estimates derived from cross-sectional data in the 1980 Census, while the second column shows panel estimates derived from relative wage *changes* within SMSAs with varying immigration flows during the 1970's.

Several aspects of these results are noteworthy. First, even the largest of the crowding effects we estimate is small. For example, as shown in panel A, a 100 percent increase in the rate of new immigration (0-5 years since entry) to the typical SMSA reduces the wages of new immigrants by only 2.4 percent. Second, the wage penalty for membership in a large immigration cohort is smaller for immigrants who have been in the United States longer. This suggests that as immigrants assimilate they "melt" into the U.S. labor market by becoming better substitutes for natives. This is

TABLE 3—THE EFFECTS ON WEEKLY WAGES OF A 100 PERCENT INCREASE IN THE SIZE OF AN IMMIGRATION COHORT

Years Since Immigration	Source	
	1980 Cross Section	1970-80 Panel
A: Own Effects on Immigrants		
1-5	-.024 (.008)	-.024 (.010)
6-10	-.026 (.008)	-.025 (.010)
11-15	-.009 (.011)	-.006 (.009)
16-20	.010 (.013)	.019 (.012)
21-30	.019 (.010)	-.007 (.010)
B: Cross Effects on Young Natives		
	Wages	Annual Earnings
Blacks	-.007 (.004)	-.004 (.006)
Hispanics	.005 (.007)	.010 (.007)

Source: Our paper (1991a, Tables 6.7 (col. 2), 6.10B, col. (2) and (4), 6.12, col. (2)), and accompanying text.

Notes: Estimates reflect the change in log earnings or weekly wages for the indicated group, relative to immigrants in an SMSA who had been in the United States 30 years or more. Effects on natives refer to the impact of a change in post-1965 immigration on black or Hispanic earnings. The characteristics controlled for are years of schooling, potential experience, marital status, number of children, disability status, race and ethnicity, occupation, and industry.

important, since it implies that immigration's effects on natives, who by definition are fully assimilated, are likely to be negligible. This point is confirmed in panel B of Table 3, where we report estimated effects of new immigration on the wages of young native black and Hispanic men, whose skills may be most substitutable for those of immigrants. These effects are smaller than the direct effects in panel A, and we regard them as economically negligible.

The foregoing estimates reflect immigration's impact on wages, leaving open the possibility that there are larger effects on employment and unemployment. One way of testing this possibility is to replicate the preceding analysis, but replacing log weekly

wages with log annual earnings (wages times weeks worked). Our 1991a paper shows that the estimates in panel A are virtually unchanged by this procedure, which implies that all of immigration's impact on earnings comes through its effect on wages; weeks of employment are unaffected. Over all, we have been unable to adduce any evidence of important distributional effects of immigration on wages and employment.

IV. Conclusion

At a time when new immigrants are entering the United States in the largest numbers in recent history, wages and employment prospects of less-skilled Americans have fallen dramatically. At the same time, new immigrants' earnings have fallen, raising questions about the long-term quality of new immigrant waves. It is tempting to see a connection between these facts and to advocate changes in policies that affect the quantity and quality of new immigrants to the United States.

Our evidence suggests that these policy concerns are exaggerated. It is true that immigrant quality, as measured by initial earnings, has declined as source countries have shifted toward Asia and Latin America. But these immigrants assimilate rapidly, and our result suggests that their long-run earning potential will be much like ethnically similar natives. Thus the long-run impact on productivity and income distribution will be much smaller than indicated by the initial earnings data. Further, although it is true that immigration has small effects

on equilibrium wages, virtually all of this burden falls on immigrants themselves. Labor market effects for nonimmigrants are negligible. Taken together, these results suggest that any adverse effects of current immigration flows on the U.S. labor market and on native welfare will be small.

REFERENCES

- Borjas, George, "Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants," *Journal of Labour Economics*, October 1985, 4, 463-89.
- _____, *Friends or Strangers: The Impact of Immigrants on the U.S. Economy*, New York: Basic Books, 1990.
- Chiswick, Barry, "The Effect of Americanization of the Earnings of Foreign-Born Men," *Journal of Political Economy*, October 1978, 86, 897-921.
- _____, "Is the New Immigration Less Skilled than the Old?," *Journal of Labor Economics*, April 1986, 4, 168-92.
- LaLonde, Robert and Topel, Robert H., (1991a) "Labor Market Adjustments to Increased Immigration," in R. Freeman, ed., *Immigration, Trade, and the Labor Market*, Chicago: University of Chicago Press, 1991.
- _____, and _____, (1991b) "The Assimilation of Immigrants in the United States: Immigrant Quality and the Changing Price of Skills," in *The Determinants and Effects of Immigration on the U.S. and Source Economies*, Chicago: University of Chicago Press, 1991, forthcoming.

THE HISTORY OF AFRICAN-AMERICAN ECONOMIC THOUGHT AND POLICY[†]

W. E. B. Du Bois and the Historical School of Economics

By THOMAS D. BOSTON*

William Edward Burghardt Du Bois was born in Great Barrington, Massachusetts, in 1868, and died in Ghana in 1963. He was a charter member of the NAACP, founder of the Pan African movement, and distinguished scholar on black life in America. Today, he is considered the father of the modern civil rights era. While most researchers are familiar with his scholarly contributions in history and sociology, few are aware of his studies in political economy.¹ Du Bois was the first African-American to pursue the Ph.D degree in economics. My objective is to acquaint the reader with this unusual individual, and to discuss his contribution to political economy and the Historical School of Economics that influenced it.

I. Historical Background

Du Bois received B.A. degrees from Fisk University in 1888 and Harvard University in 1890. Upon graduation, he entered Harvard's Ph.D program in history and political economy. His last two years of graduate study, 1892-94, were spent in residence at the University of Berlin. During this period,

Germany attracted many of America's leading graduate students. But, more importantly, the Political Economy Department at the University of Berlin was at the center of the famous "Methodenstreit," or Battle of Methods.

By 1880, the Historical School of Economics had reached its zenith under the influence of Gustav von Schmoller, J. Spiethoff, W. Sombert, A. Wagner, and M. Weber. Then in 1883, Carl Menger published a broadside assault on this school's methodology and ignited the "Methodenstreit." This battle pitted economists who employed the inductive, interdisciplinary method of research against the marginalist theoreticians led by Menger, Walras, and Jevons.

Du Bois entered the Department of Political Economy at Berlin during the high point of the Methodenstreit and his graduate supervisor was none other than Schmoller, the intellectual leader of the Historical School. His course work included theoretical political economy and industrialism and society from Schmoller and Wagner. Additionally, he sat through lectures by Max Weber and enrolled in Schmoller's research seminar.

These three professors had a profound influence upon Du Bois. In fact, Francis Broderick (1971) credits Schmoller with drawing Du Bois away from history and into political economy, and inspiring him to devote his career to scholarly research. Several years earlier, the faculty of Fisk University had nurtured Du Bois' lifelong commitment to resolving the "race problem." Now, the University of Berlin offered him the necessary methodological tools.

While studying at Berlin, Du Bois prepared a research paper that was to be pub-

[†]*Discussants:* William A. Darity, Jr., University of North Carolina-Chapel Hill; Bernadette Chachere, University of California-San Francisco.

*Associate Professor of Economics, School of Economics, Georgia Institute of Technology, Atlanta, GA 40332.

¹Herbert Aptheker (1980), Du Bois' biographer, has compiled a bibliography of his scholarly articles, essays and editorials that consumes several hundred pages; included among these are 21 books and 16 edited volumes. Du Bois' publications in political economy are concentrated between the years of 1896 and 1906.

lished in Schmoller's *Yearbook*.² Additionally, Schmoller took the unusual step of petitioning the university to admit Du Bois to Ph.D candidacy after only three semesters of study, one-half the minimum time required. In his *Autobiography*, Du Bois observed:

Schmoller wanted to present me for my doctorate, despite the fact that I had not finished the "triennium" required in a German university and my work at Harvard was not recognized. The faculty was willing in my case but was restrained by the professor of English who threatened to push the similar claims of several Britishers. I therefore regretfully had to forego the chance of a German doctorate and wait for the degree from Harvard. [1968a, p. 175]

More than one-half century later, Du Bois noted that the Berlin teachers gave him an understanding of the relation of politics and economics and the connection between European development, America's race problem, and the problems of Africa and Asia (1968a, p. 163).

II. Du Bois' Contributions to Political Economy

Upon returning to America, Du Bois completed his dissertation that became the first volume of the Harvard Historical Series (Du Bois, 1896). Additionally, he became that university's first African-American doctoral graduate.³ Three years later, he published *The Philadelphia Negro* (1873/1899) that many consider to be his finest scholarly achievement.

Between 1898 and 1906, Du Bois conducted five agricultural studies for the U.S.

²The paper, "The Plantation and Peasant Proprietorship System of Agriculture in the Southern United States," was the first of a series of studies on southern agriculture conducted by Du Bois.

³His degree was awarded in history, but his dissertation and studies were supervised by the departments of History and Political Economy.

Bureau of Labor that clearly exhibit the influence of the Historical School.⁴ Some of his other publications in political economy during this period include a paper prepared for the American Economic Association and one solicited by Max Weber for the journal he edited (see Du Bois, 1906a, b).

Unfortunately, the study that Du Bois considers to be his most significant was destroyed by the Bureau of Labor. He recounts that the report was finished by hand with no copy and rushed to Washington where he was told that it would not be published because it "touched on political matters.... I went back to them again to ask if they would allow me to have the manuscript since they were not going to use it. They told me it had been destroyed!" (1968a, p. 227).

Between 1897 and 1910, Du Bois accepted an appointment as Professor of History and Political Economy at Atlanta University. During this period, he initiated an ambitious research agenda for the study of black life. Its success, however, was contingent upon adequate financial support and the cooperation of several leading majority institutions. With adequate support, he argued, this effort would culminate in one of the greatest scholarly endeavors ever undertaken. But the support never materialized. Still, he edited sixteen volumes of research on black life (1968b).

III. The Historical Method

Du Bois' research in political economy is clearly influenced by the historical method. This method rejects abstract theoretical analyses and argues instead that economic life must be studied concretely. Truth, Schmoller proclaims, is only revealed through detailed historical analysis of descriptive material supplemented by a variety

⁴Included among these publications are "The Negro of Farmville Virginia" (1898), "The Negro in the Black Belt" (1899), "The Negro Landholder in Georgia" (1901), and "The Negro Farmer" (1906). (They are collected in Aptheker, 1980.)

of statistical investigations (Karl Pribram, 1983). The method also emphasized the interdependence of economics and other aspects of social life.

The Historical School repudiated Classical Economics on the premise that it erred in making excessive deductions from erroneous empirical facts. While occasionally admitting that economic laws exist, historicists argued that the most effective means of uncovering them is through historical inquiry. Economists must have a "hunger for facts," they asserted, and a willingness to examine the general body of history in great detail. Only then can they derive patterns, processes and generalizations.

Not surprisingly, economic theory suffered under the influence of the historicists. In fact, Schumpeter calls this school's theory "pedestrian." But he credits it with producing unrivaled historical studies that caused a tremendous advance in the accuracy and knowledge of social processes.

Finally, the historicists were extremely active in social reform movements. Yet their scientific credo criticized the injection of value judgments into research. "There was respect for the economic fact," Schumpeter asserts, "and the will to let it speak for itself" (1954, p. 811).

IV. The Historical School's Influence on Du Bois

The extent of descriptive and statistical detail in Du Bois' studies is rarely matched even today. Further, the lack of emphasis on theory strongly resembles the historical method. Speaking of the *Philadelphia Negro*, he asserts, "Of the theory back of the plan of this study of Negroes I neither knew nor cared. I saw only here a chance to study a historical group of black folk and to show exactly what their place was in the community" (1968a, p. 197).

Du Bois' approach matched the Historical School's in two other respects. First, he always studied economic events in the context of their surrounding social phenomena. Second, he never injected normative judgments into scholarly research. Although he was a social activist throughout his life, he

confined his propaganda for racial equality to journalistic editorials and speeches.

Like other economists of the Historical School, Du Bois objected to abstractions and normative judgments in scholarly research. "We simply collect the facts" he once observed, "others may use them as they will." His only predilections were: 1) that African-Americans are part of the human race and have the same endowment and capacity for culture and development as all others; and 2) that African-Americans are deserving of as much scientific investigation as are all other ethnic groups.

V. Du Bois, a Little-Known Economist

Today, if an economist is asked to identify the definitive study of black life, he will inevitably answer Gunnar Myrdal's *American Dilemma*. Yet Myrdal has stated that this status belongs to Du Bois' *The Philadelphia Negro*. If this is the case, then why is Du Bois so unfamiliar to most economists? I close with a consideration of this issue.

Du Bois' contribution to political economy is partly obscured by the historical method that influenced it. The Historical School withered under the assault of the Marginalist Revolution and, by 1910, its method was relegated to economic historians and sociologists. Accompanying its demise, the study of economics became increasingly theoretical and more isolated from other social sciences.

During this same period, Du Bois turned his attention more and more to the study of sociology, a discipline that separated formally from the American Economic Association in 1905.⁵ These occurrences serve to distance his scholarly research from mainstream economics. They also help explain why so few economists are aware of his contributions in political economy. Today,

⁵The research of Max Weber, one of Du Bois' former instructors, also moved in this same direction and today he is noted more as a sociologist than an economist.

his publications are not thought of as economic research. But, at the turn of the century, they were contributions to mainstream historical economics.

There is a second and perhaps more important circumstance that obscures Du Bois' contribution to political economy. Although his research was recognized internationally, as an African-American he was never offered an appointment at a predominantly white university. Confined by race to teach at historically black colleges, he still appealed to the leading academies to conduct research on black life jointly with him. But his appeals were rejected. He notes that the only post he was ever offered at a major white university was a one-year position in the "unusual status of 'assistant' Instructor" (1968a, p. 194).

The constraints that racial inequality imposed upon his research were aggravated further by his famous political debate with Booker T. Washington. Unfortunately, the latter's political influence with white benefactors of Atlanta University led to Du Bois' resignation in 1910. Thus, his most fertile period of scholarly productivity ended and he embarked upon a career of social activism.

Disenchanted with racism and capitalism and falsely accused of being an agent of a foreign government by the McCarthy inquisitions of the 1950's, Du Bois went into self-imposed exile in Ghana in 1961. Two years later, he died while directing the *Encyclopedia Africana* project.

REFERENCES

- Aptheker, H., *Contributions by W. E. B. Du Bois in Government Publications and Proceedings*, Millwood: Kraus-Thompson, 1980.
- Broderick, Francis L., "The Search for a Career," in Rayford Logan, ed., *W. E. B. Du Bois: A Profile*, New York: Hill and Wang, 1971, 1-37.
- Du Bois, W. E. B., *The Philadelphia Negro*, Millwood: Kraus-Thompson, 1973/1899.
- _____, (1968a) *The Autobiography of W. E. B. Du Bois*, New York: International, 1968.
- _____, (1968b) *Atlanta University Publications*, 2 Vols., New York: Octagon Books, 1968.
- _____, (1906a) "The Economic Future of the Negro," *Publications of the American Economic Association*, February 1906, 7 (3rd. ser.), 219-42.
- _____, (1906b) "The Negro Question in the U.S.," *Archiv Fur Sozialwissenschaft und Sozialpolitik*, Tübingen, 1906, 22, 31-79.
- _____, *The Suppression of the African Slave Trade to the U.S.*, Harvard Historical Series, No. 1, New York: Longmans, Green, 1896.
- Pribram, Karl, *A History of Economic Reasoning*, Baltimore: Johns Hopkins University Press, 1983.
- Schumpeter, Joseph, *History of Economic Analysis*, New York: Oxford University Press, 1954.

Missed Opportunity: Sadie Tanner Mossell Alexander and the Economics Profession

By JULIANNE MALVEAUX*

Sadie Tanner Mossell Alexander (1898–1989) was the first black woman in the United States to receive a Ph.D. (Thomas Potterfield, 1990). She earned her degree in economics from the University of Pennsylvania in 1921. Her dissertation, “The Standard of Living Among One Hundred Negro Migrant Families in Philadelphia,” was, in her words “an attempt to arrive at conclusions concerning the migrants to Philadelphia, through an intensive analysis of the budgets of a small number of their group” (Mossell, 1921). When she finished her graduate work in 1921, Mossell was unable to find employment in the economics profession in Philadelphia or in the surrounding areas (Potterfield). This is not surprising, since the only academic employment available to African-Americans in the early 1920’s was available at the black colleges, and black women faced barriers to employment at some of those colleges.

In any case, Mossell worked as an assistant actuary at the black-owned North Carolina Mutual Life Insurance Company from 1921 to 1923. She returned to Philadelphia to marry Raymond Alexander in 1923, and entered the University of Pennsylvania Law School in 1924. She was the first African-American woman to graduate from Penn’s Law Schools in 1927, and the first black woman admitted to the Pennsylvania bar in the same year (Potterfield).

Given her family background, Alexander’s academic and career achievements are not surprising. Her grandfather was a bishop in the African Methodist Episcopal Church. Her uncle, Henry O. Tanner, was a noted artist. Nontraditional occupations and employment were not unusual for the Tanner women—an aunt, Hallie Tanner Johnson,

was a physician and founder of the nurse’s school and hospital at Tuskegee Institute (Potterfield).

During her career, Alexander distinguished herself as a lawyer in both the public and private sectors, working as Assistant City Solicitor in the City of Philadelphia in 1928–30 and 1930–34, and serving as a Truman appointee to the Committee on Human Rights (Gerald Fraser, 1989), a Kennedy appointee to the Lawyer’s Committee on Civil Rights, and the Carter-appointed Chair of the White House Conference on Aging (Fraser). She was also active in civic affairs, serving as the Secretary of the National Urban League for 25 years, and as a member of the National Advisory Council of the American Civil Liberties Union. She was the first national president of Delta Sigma Theta Sorority (currently the largest black woman’s public service organization; see Paula Giddings, 1988), and its legal advisor for 35 years.

Alexander’s household seems similar to the modern two-career household. Her husband was a lawyer, a member of the Philadelphia City Council (1951–59), and a distinguished jurist on the Philadelphia Court of Common Pleas. They had two daughters, Mary Elizabeth Alexander Brown (born 1934) and Rae Pace Alexander-Minter (born 1936). Alexander successfully juggled household, family, career, and civic responsibilities. The achievements of her life touch areas that range from the law, government, civil rights, education, aging, and women’s rights (Potterfield).

A striking aspect of Alexander’s career was the longevity of her interest in the topics she identified as important—serving as Secretary of the Urban League for 25 years, and as the Legal Advisor to Delta Sigma Theta for 35 years. Her involvement in the economics profession, on the other hand, seems fleeting. Alexander earned a Ph.D. degree, sought employment in economics

*Visiting faculty, Afro-American Studies Department, UC-Berkeley, CA 94720.

but did not find it, and worked as an actuary for a short time before pursuing legal studies. The Alexander papers suggest that Alexander all but abandoned the economics profession after 1923,¹ although at least one of the books in the Alexander collection suggests that 30 years after obtaining the Ph.D., she maintained at least a passing interest in economic issues.²

Why did Alexander abandon economics? According to her daughter, Rae Alexander-Minter, "there was no way for her to make a living in the profession." Alexander-Minter indicates that her mother did not look back and expressed no regrets about her career, but worked hard at her law practice and at maintaining her family. Indeed, at her peak, Alexander had one of the largest divorce practices in Philadelphia.³

While her withdrawal from the economics profession may not have been a personal tragedy for Alexander, the fact that she did not continue her work in economics seems a missed opportunity for her, for the economics profession, and for the body of economic knowledge that pertains to African-Americans. My goal in this essay is to probe the nature of that missed opportunity, both through discussion of the major example of her economics work, the doctoral dissertation, and through speculation about ways she may have followed up on the dissertation, given her interests.

I. Black Migration and Consumption in Philadelphia, 1916-18

According to Philip Foner and Ronald Lewis (1989), the "machinery of segregation" had been installed in the South by the beginning of the twentieth century. "Economic intimidation, violence, and lynching"

suggested to blacks that the South had no future for them. In addition, the early twentieth century was an economically devastating period in the South, with agriculture plagued by flooding, an epidemic of boll weevils, and other hardships. As the South declined, the North and Middle West developed industrially, especially with demand stimulated by the production needs of World War I. As white men went to war, black men moved North to replace them in industry. "The Great Migration," a significant exodus of black workers from the South, took place in the second decade of the twentieth century, triggered both by hardship in the South and by new opportunities in the North.

Between 1890 and 1915 in Chicago, for example, the black population grew from less than 15,000 to more than 50,000 (Allan Spear, 1967). Similar population jumps took place in other industrial centers, including Philadelphia. Alexander's dissertation, "The Standard of Living of Negro Families in Philadelphia" (henceforth referred to as NMF), looks in detail at this migration in Philadelphia, both from a macro and a micro perspective.

Alexander documents the labor shortage in Philadelphia, and discussed industrial efforts to attract black workers to manufacturing sites. NMF discusses the living conditions for migrant families, conditions Alexander describes as deplorable. She goes on to detail the civic response to poor housing conditions, on the part of the black church, the black middle class, and whites. Some of the racial conflicts that took place because of migration are detailed by Alexander, mainly to place the issue of migration into a sociopolitical context, and to develop research questions and pose hypotheses for her dissertation.

"Was the migrant to Philadelphia able to adapt himself to the environment of an industrial economy?" she asks. "Did his presence help or hinder the racial condition of the city?" Alexander asserts that the standard of living maintained by a people is an index of the extent to which they have adapted themselves to a given environment. She proposes to analyze the incomes and expenditures of a group of migrant families

¹Telephone conversation between Mark Lloyd, Director of University of Pennsylvania Archives and Records center and myself, September 21, 1990.

²According to Potterfield (p. 229): Among the 22 miscellaneous books included in the joint papers of Raymond Pace Alexander and Sadie T. M. Alexander was *Away From Freedom: The Revolt of the College Economists*, edited by Orval Watts in 1952.

³Personal interview with Rae Alexander-Minter, December 10, 1990.

to measure standards of living and "judge the degree of adaptation."

Thus, Alexander set the tone for doing a cross-sectional consumption survey of black migrant families. Over a 2-month period, Alexander visited 100 families in Philadelphia's twenty-ninth ward. Using a detailed questionnaire, she asked about origin, family structure, labor market status, unemployment, and household expenses, including rent, utilities, church, insurance, tobacco, alcohol, carfare, and savings. In a manner reminiscent of contemporary dissertations, Alexander discusses both her research methodology and the accuracy of her data before analyzing results.

The results of Alexander's dissertation provide fascinating information about expenditure patterns, and also reveal her familiarity with federal and municipal data on household incomes and consumption patterns, and theories of consumption and spending. NMF also reveals Alexander's penchant for detail and accuracy, as well as a good eye for reporting of the minute aspects of migrant lives. Discussing expenditures on heat, for example, Alexander expanded the discussion to deal with families who rented rooms where landlords provided heat and ventilation and health aspects of these rentals. She concluded that many of these dwellings "were unfit for families."

NMF reports many of the economic inefficiencies that migrant families were forced to endure. "The average price paid for renting one room was \$163 per year, \$6.05 less than average price of renting a house of four rooms," writes Alexander. She adds that renting a home increased income earning potential since so many renters took in lodgers. Noticeably absent from this discussion, however, was an acknowledgement of the institutional forces that forced black migrant families into inefficient economic arrangements.

Alexander's discussion of migrant spending provided information about the culture and lifestyles of black migrants. She details information about health and the use of free clinics, the financial ties among extended families, and savings patterns, including participation in benevolent societies and church thrift clubs. Alexander con-

cludes the discussion of expenditures by comparing black migrant expenditure patterns with those posited by consumption theorists and those measured by the war labor board. Essentially, she found that black migrant families behaved in the same way Engel theorized they would. The percentage of income spent on food fell as income increased. The percentage of income spent on clothing stayed the same with increasing income. The percentage of income spent on sundries increased as income increased. In a result that differed from Engel's theory, expenditures on rent, fuel, and light decreased as income rose.

After analyzing black migrant budgets and expenditures, Alexander developed a suggested budget. In the context of this discussion, especially as it relates to housing, she dealt with supply and segregation issues and indicated that "a two story brick house, fitted with tub, washstand, and toilet" might be scarce for blacks because "the Negro population of Philadelphia increased without an equal increase in housing." Alexander put off issues of housing availability for the concluding chapter of her dissertation.

Armed with data about actual and suggested spending, Alexander asks two questions. "Do black migrant families earn enough to have fair standards of living, as defined by published reports?" "Do black migrant families choose to spend their money in ways to attain a fair standard of living?" She notes that 64 percent of all families had a sufficient income to provide a fair standard of living, though the primary breadwinner's earnings were enough to provide a fair standard in only 41 percent of cases. She further indicates that some of the families able to afford a fair standard of living do not attain such standards because of "unwise" spending.

In this context, Alexander deals with issues of class and migration in distinguishing segments of the black community from each other. She notes that family size, "ignorance resulting from unwise spending" (insurance spending is especially targeted here), and racial prejudice were all factors preventing black migrant families from attaining a fair standard of living. She deals with both institutional barriers to fair living standards

("Recreation appeared seldom in his budget, for the Negro was admitted to few places where it was offered"); and individual barriers (such as lack of education). Alexander concludes by suggesting that the status of the black migrant can be improving by the "Negro businessman," the black church, and the city, which she says "has the responsibility of seeing that at least adequate housing is secured."

II. Missed Opportunity: Beyond Negro Migrant Families

Alexander's doctoral dissertation illustrates the perception, sensitivity, racial concern, and ability of a young upper middle-class black woman to use her professional skills to tackle a contemporary racial issue. Her dissertation is a case study that reveals institutional aspects of racial segregation in Philadelphia, as well as confirms the consumption theories of the time. It also reveals Alexander's middle-class biases, for example, in her use of terms like "unwise purchases" and her remark that she does not advocate the consumption of alcohol, Prohibition or not. On the other hand, in her concluding chapter Alexander reveals herself both as an advocate for black self-help ("churches could help to alleviate the housing problem by building houses instead of expensive church edifices") and as an advocate of government economic involvement in the housing market.

In thinking of the missed opportunity revealed by Alexander's dissertation, I think about the work that Abram Harris did on black business development, Oliver Cox's work on class, the W.E.B. DuBois Atlanta University studies, and D. Parke Gibson's books on black consumer patterns. Given her interest and ability, Alexander might have followed in any of those directions or, indeed, continued to look at black family and migration patterns both through further case studies and from a macroeconomic perspective. Because Alexander discussed so many possible areas for further research in her dissertation, one can posit that, given the opportunity, she would have had a productive and significant research career.

Indeed, it is possible to speculate that had Alexander pursued her research interests, given her long involvement with the Urban League, we might have seen the earlier production of titles like "The State of Black America," which was first produced in 1975 (John Jacob, 1991). And, given the opportunity to teach, what kinds of students might Alexander have nurtured, and how many other economists would have tackled socioeconomic issues of migration, consumption, housing availability, health care access, and insurance availability?

Alexander's dissertation suggests that the young economist might have taken quite a different career path after she earned her Ph.D. had there only been opportunities for her in her profession of first choice. While the economics profession's loss was the legal profession's gain, Alexander might well have made a significant contribution to economics, given the opportunity.

REFERENCES

- Alexander, Sadie Tanner Mossell, *Vitae*, 1982.
- Foner, Philip S. and Lewis, Ronald L., *Black Workers; A Documentary History from Colonial Times to the Present*, Philadelphia: Temple University Press, 1989.
- Fraser, Gerald, "Sadie T. M. Alexander, 91, Dies; Lawyer and Civil Rights Advocate," *New York Times*, November 3, 1989.
- Giddings, Paula, *In Search of Sisterhood: Delta Sigma Theta and the Challenge of the Black Sorority Movement*, New York: William Morrow, 1988.
- Jacob, John, "Urban League Releases State of Black America 1991 Report," *Urban league News*, January 8, 1991.
- Mossell, Sadie Tanner, "The Standard of Living Among One Hundred Negro Migrant Families in Philadelphia," in *Annals of the American Academy of Social and Political Science*, November 1921.
- Potterfield, Thomas G., "Guide to the Alexander Papers," University of Pennsylvania Archives and Records Center, 1990.
- Spear, Allan, *Black Chicago: The Making of A Negro Ghetto*. Chicago: University of Chicago Press, 1967.

The Rise and Fall of Negro Economics: The Economic Thought of George Edmund Haynes

By JAMES B. STEWART*

This analysis examines the treatment of economic issues in studies by George Edmund Haynes. Born in 1880, Haynes became one of the first African-Americans to earn a Ph.D. in sociology. He was awarded the doctorate by Columbia in 1910 after earlier periods of study at Fisk, Yale, and the University of Chicago. Haynes' formal connection with the economics profession derives in part from his tenure as Director of Negro Economics for the U.S. Department of Labor during World War I. That appointment resulted primarily from the visibility of Haynes' work with the Urban League.

Haynes was serving as the League's Educational Secretary and was Professor of Social Science at Fisk University at the time of his selection for the Negro Economics directorship. Earlier he had completed an investigation of the adjustment of black migrants to New York as well as a general study of the condition of blacks in cities (see his 1912, 1913 articles). Coincident with his tenure as Director of Negro Economics, his case study of the adaptation of black migrants in Detroit was published in 1918.

Haynes' appointment to a position labelled "Director of Negro Economics" was somewhat paradoxical. The paradox derives from the underlying theme of his work, namely similarities in the effects of economic and social forces on blacks and whites. This focus led to the understatement of the role of discrimination in creating barriers to the achievement of equality of condition.

Many of Haynes' perspectives anticipated contemporary economic approaches to the study of social problems. Although he lacked

access to modern economic categories, the contemporary labels appropriate for characterizing the foci of Haynes' research include the economics of rural-urban migration, the structure and functioning of urban labor markets, economic discrimination, and the underground economy. Haynes approached these areas with the eye of a social change agent rather than as a theoretician. But despite this practical problem-solving orientation, his investigations were undergirded by economic models of individual behavior that should be familiar to modern economists.

The similarities between Haynes' perspectives on the economic circumstances of African-Americans and much contemporary scholarship underscore the usefulness of a detailed examination of his approaches to the study of each of the subject areas mentioned previously.

I. The Economics of Migration

Haynes (1918) employed a general push-pull model to analyze forces contributing to the northern migration of blacks in the early twentieth century. For Haynes, "The cause first in importance...is economic" (p. 6). He argued that between 1914 and 1918, the districts that had "the worst economic conditions to face...have lost the largest numbers of their Negro population" (p. 6). He argued that an extraordinary demand for labor had been created by World War I and labor shortages were exacerbated by the disruption of European emigration. As a result, northern employers discovered "the unworked Negro labor supply of the South" (p. 6).

Haynes suggested that the size of the black migration was in part the result of a "money illusion." In his words, blacks were "thinking in terms of money wages not real wages" (1918, p. 7). This money illusion was

*Department of Labor Studies and Industrial Relations, Pennsylvania State University, University Park, PA 16802.

promulgated by tales of high wages told by northern labor agents. He also noted the use of free transportation as an incentive to migrate.

Although Haynes saw economic incentives as paramount, the "social causes of migration" were also described. These included poor housing, poor schools, Jim Crow segregation, disfranchisement, discrimination in the criminal justice system, and mob violence. Haynes also discussed the role of public policy in determining the configuration of the opportunity set of potential migrants. Different public thrusts in the North and South had made the North a relatively more attractive region of residence. In the North, public education and public goods such as parks and recreational centers were identified as important initiatives. In the South, public efforts to stem the exodus of blacks through regulation of northern labor agents and tightening debt peonage provisions were highlighted as indicators of increased levels of oppression.

Haynes maintained that his model of migration could be applied to other periods and to other population groups. The post-1910 migration was for him, "only the third swell in a movement...which ha[d] been in progress since 1860" (1918, p. 6). In the New York study he asserted that, "the general movement of the Negroes...does not seem to have been very different from that of whites" (1912, p. 14). In fact, the movement of the white and black populations towards cities was hypothesized to be coincidental.

The obvious implication of the view that underlying causes of migration were similar across populations was that "the problems which grow out of his [blacks] maladjustment to the new urban environment are solvable by methods similar to those that help other elements of the population" (Haynes, 1912, p. 14).

II. The Structure and Functioning of Urban Labor Markets

In modern parlance, Haynes employed a segmented labor market model to analyze the structure and functioning of urban labor

markets. The primary sector consisted of jobs in the industrial sector while the secondary sector was dominated by service occupations. Changes in aggregate demand and variation in labor market conditions were treated as potential sources of reductions in inequalities generated by segmentation.

In the New York City study, Haynes found that most blacks were employed in domestic and personal service occupations (1912, p. 77). However, for Detroit, he observed a much wider range of employment of African-Americans (1918, p. 15). One of the obvious sources of the differences was the demand for labor created by World War I. In fact, Haynes observed that previously blacks in Detroit had been "confined to some of the domestic and personal service occupations [and h]e was probably losing ground in these until the industrial demand of the Great War came" (1918, p. 12).

The effects of blacks' relegation to the secondary sector was viewed as producing a cycle of diminished economic well-being such that "there is a larger number of competitors within a limited field with a consequent tendency to lower an already low wage scale. In this way the limitations of occupational mobility react upon income, producing a low standard of living, the lodger evil, and social consequences..." (Haynes, 1912, p. 77).

Haynes was especially concerned about mothers of young children in the labor market and crime. On the issue of crime he observed, "Arrests and prison commitments have many factors which figures do not show and are quite as much a commentary upon the white communities at large as upon the unfortunate Negro law breakers" (1912, p. 39). The "lodger evil" was the practice of renting rooms in the home to supplement family income. Haynes concluded that many women had to "help earn the daily bread of the family" and that "Their low income power forces these families to the necessity of completing the rent by means of lodgers, deprives the children of mothers' care, keeps the standard of living at a minimum, and thus makes the family unable to protect itself from both physical and moral disease"

(1912, p. 89). Haynes found that in New York approximately half of the households included lodgers.

Three factors were hypothesized by Haynes to account for the relegation of blacks to the secondary sector: (a) "historical conditions of servitude"; (b) "a prejudice on the part of white workmen and employers"; and (c) "inefficiency of Negro wage-earners for competition in occupations requiring a higher order of training and skill" (1912, p. 77). Most of Haynes' attention was focused on the "inefficiency" problem despite his observation that the age distribution of black workers in New York City suggested that "They are either killed off by the conditions under which they work and live, or drift away from the city at a premature old age" (1912, p. 57).

Haynes seemed to take at face value Detroit employers' assertion that many black workers did not work fast enough and were "disinclined to work out-of-doors when the cold weather comes" (1918, pp. 17-18). He recommended that the churches "provide lectures, social study classes to overcome these faults" (p. 20). This instruction was to include "instruction in how to dress and in simple repeated assurances that [workers] need not fear freezing to death if they are properly clad" (p. 20). Haynes also proposed training programs for domestic workers as a major initiative. This proposal was curious because, he had argued elsewhere that "the large majority of Negroes in domestic and personal service are capable, temperate, and honest, and remain with one employer a reasonable time..." (1912, p. 89).

Haynes saw investments in human capital as critical both for gaining access to occupations opened up by labor shortages and for retaining those jobs. His recommended interventions emphasized broad involvement of voluntary organizations like the Urban League and churches to change the characteristics of workers. He focused little attention on exploring ways to reduce segmentation. One of his most fascinating discussions of this issue was associated with an experiment involving a pants manufacturing plant employing only black women (excluding

management). Haynes saw this experiment as contributing to black workers gaining a permanent hold in Detroit industries (1918, p. 19). Contemporary perspectives would, of course, contest this assessment. To understand Haynes' views requires discussion of his theory of economic discrimination.

III. The Economics of Discrimination

Haynes' limited discussions of discrimination are carefully worded and strategically located in the various analyses. It is the very last statement in the New York City study where the declaration appears that "Negro wage-earners and business men have great difficulty in scaling the walls of inefficiency and of race prejudice in order to escape the discomforts and dangers of a low standard of living" (1912, p. 148).

The principal theory of discrimination employed by Haynes was a forerunner to modern statistical theories of discrimination. He argued that one of the consequences of the large-scale migration of unskilled blacks to urban areas was an increase in residential segregation. This segregation limited the contact between blacks and whites leading to the assessment of the capabilities of individuals on the basis of ascribed group characteristics. In Haynes' words,

[T]oo much of the white man's knowledge of the Negro people is derived...from domestic servants and from superficial observation of the loafers about the streets. The best elements of both races, thus entirely removed from friendly contact, except for the chance meeting in the marketplace, know hardly anything of their common life and tend to become more suspicious and hostile toward each other than toward strangers from a far country. [1913, pp. 110-11]

Haynes maintained that "prejudice, when displayed by employers, is partly due to inefficiency...and the failure to discriminate between the efficient individual and this untrained throng" (1913, p. 112).

Haynes also allowed, however, that in some cases white employers and employees simply had a Becker-type "taste to discriminate," that is, "in the cases of many employers and employees...the opposition to the Negro in industrial pursuits is due to a whimsical dislike of any workman who is not white and especially of one who is black!" (1913, pp. 112-13). The statistical model of discrimination was used to assign blame to the victim for fueling discriminatory sentiments while the Becker-type model mildly chastised the discriminator. On balance his recommendations placed the principal burden of adjustment on the victim. He proposed that whites accord blacks equal justice via "a square deal in industry, in education, and in other parts of the common life" (1913, p. 118). However, his first recommendation was as "an organized effort to acquaint the Negro in the country with the desirability of remaining where he is unless by education and training he is prepared to meet the exactions of adjustment to city life" (1913, pp. 118-19). Thus for Haynes, it was better for blacks to remain in the shadow of the plantation than to risk aggravating aggregate adjustment problems in the North.

IV. Conclusion

We can only speculate as to what recommendations Haynes would have for residents in contemporary urban environments who cannot "meet the exactions of adjustment to city life." However, constructions similar to those of Haynes can be found in many contemporary neoconservative analyses (Thomas Sowell, 1981; Walter Williams, 1982). Neoconservative analysts tend, however, to employ much less sophisticated analytical frameworks than Haynes.

Haynes applied his basic approach to other issues of contemporary interest including black business enterprises (1912). His study of the involvement of adolescents

in junk dealing could be applied easily to examine the contemporary underground economy (Harry Grigg and Haynes, 1919). Haynes discussion (1950) of Africa as the continent of the future extended the application of many of his models to the international arena. "In the field of social values," he wrote, "the Africans hold still to kinship and family ties and communal interdependence that may offer valuable patterns to a world writhing in rugged individualism" (p. 20). Haynes offered a scenario that may have important implications for the next century; "European and American workers may well ponder what will happen when factories...begin to ship their products made with cheaper wages and raw materials, to the world markets in vessels built in African shipyards" (1950, pp. 20-21).

REFERENCES

- Grigg, Harry and Haynes, George, *Junk Dealing and Juvenile Delinquency. An Investigation Made for the Juvenile Protective Association of Chicago*, Chicago: Juvenile Protective Association of Chicago, 1919.
- Haynes, George E., *Africa: Continent of the Future*, New York: Association Press, 1950.
- _____, "Conditions Among Negroes in the Cities," *Annals of the American Academy of Political and Social Science*, September 1913, 49, 105-19.
- _____, "The Negro at Work in New York City, A Study in Economic Progress," in *Studies in History, Economics and Public Law*, 1912, Vol. 49, No. 3, New York: Longman, Green and Co., 1912.
- _____, *Negro New-Comers in Detroit, Michigan, A Challenge to Christian Statesmanship: A Preliminary Survey*, New York: Home Missions Council, 1918/1969.
- Sowell, Thomas, *Markets and Minorities*, New York: Basic Books, 1981.
- Williams, Walter, *The State Against Blacks*, New York: New Press, 1982.

Celestial Mechanics and the Location Theory of William H. Dean, Jr., 1930–52

By JULIAN ELLISON*

William Henry Dean, Jr. was born in 1910 and died in 1952. He was the only son, and third of four children, of Reverend William Henry Dean and the former Ella Cornelia Green. His father was a Methodist minister, and the family lived in the cities in which he pastored churches: Lynchburg, VA; Washington, D.C.; Baltimore, MD; and Pittsburgh, PA. The son graduated from Frederick Douglass High School in Baltimore in 1926 as valedictorian of his class. He attended Bowdoin College in Maine, graduating summa cum laude in 1930. He then entered Harvard University, where he earned the M.A. and the Ph.D. in economics in 1932 and 1938, respectively. His subsequent career included stays with Atlanta University, 1933–42; City College of New York, summer 1939; U.S. National Resources Planning Board, 1940–42; U.S. Office of Price Administration, Virgin Islands, 1942–44; National Urban League, 1944–46; and United Nations, 1946–52 (Rayford Logan and Michael Winston, 1982).

This paper examines briefly the sources of location theory in the mathematics and astronomy of the day, and Dean's application of results from these fields to economic location theory in his doctoral dissertation at Harvard University in 1938.

I. The Development of Economic Location Theory Before Dean

A line of influence stretching from Joseph-Louis Lagrange (1772) to Georg A. Pick (Weber, 1929) to George D. Birkhoff (1915) to Halford J. Mackinder (1902) to

Dean (1938) accounts for the peculiar contours of Dean's economic location theory. These names represent simply the peaks observable from the present on an azimuth from Dean to the initial modern application of geometry to celestial space. It should not be presumed, and is here expressly disclaimed, that this is an exhaustive account of the process of development involved in the Dean accomplishment. In particular, only allusions are made to the very important roles played by city and regional planning, geography, international trade theory, and by the *Quarterly Journal of Economics* and the *Review of Economic Statistics* in the creation of Harvard location theory (Sidney McCluskey, 1963; Erwin Finlay-Freundlich, 1958; Otto Dziobek, 1962; Victor Szbeheley, 1967; Eric Bell, 1965; Birkhoff).

II. Dean's Doctoral Dissertation and Its Contribution to Location Theory

Dean's entire original dissertation was comprised of 8 chapters divided into two parts: Part I, "Pure Theory of Industrial Location," consisted of the first 3 chapters; Part II was entitled "Location in the Geographic and Historical Environment," and contained the remaining chapters. His published book in 1938 contained Part I and chapter 5.

The idea of a plane triangle or polygon as the representation of the industrial location problem was taken from the restricted problem of three bodies in celestial mechanics. Dean's book contained three sections illustrating the relationship of location theory to celestial mechanics. In chapter 2, "The Regional Pattern: Static Theory of Industrial Location," he constructed a locational triangle and a weight triangle, that were interpreted as showing the relationships among sites of input sources, production, and market. Each of these sites represented mass. Their positions were determined by analogy

*Mid-Atlantic Research Corporation, 9318 Grazing Terrace, Gaithersburg, MD 20879. A complete list of references for all articles and books cited herein may be obtained from the author upon request.

to the laws of gravity and celestial motion of Isaac Newton (1946), Lagrange, Poincaré (1957, 1905–10, 1912), and Birkhoff, and in particular by analogy to the restricted problem of three bodies.

In his article, Birkhoff presented a geometrical diagram representing the restricted problem. The diagram showed an isosceles triangle superimposed on a rectangular coordinate system, with the vertices of the triangle representing the points of mass of the three bodies, and the sides representing the motion vectors. Dean used a virtually identical diagram to illustrate the locational triangle.

In chapter 3, "The Regional Pattern: Dynamic Strategy of Industrial Location," Dean alluded to a second great force of nature, electromagnetism, in formulating his theory of location. Dean here followed James Jeans (1908) in his analysis of electricity, and identified electrical power as one of the two major technical determinants of location, along with surplus food.

In chapter 5, "Nodality and Commercial Agglomeration. Selections," Dean returned to astronomy to find the concept of a node, or the intersection of the orbits of celestial bodies. The node concept had been applied to geography by Mackinder, who considered that nodes in human geography had positive consequences. Dean, indeed, argued that geographical nodes, or intersections of human trade routes, are likely solution sites for the location of industrial production facilities. Moreover, since such nodes develop at topographically suitable sites, Dean here abandoned the featureless plain preferred by Carl Launhardt (1872), Alfred Weber/Georg Pick, Frank Fetter (1924, 1931), Harold Hotelling (1929), and Edgar Hoover (1937) as the environment of their location theories, and moved in a more "realistic" direction.

Dean's most important contribution, therefore, was to apply the mathematical models and methods of analysis of celestial mechanics to the solution of spatial problems in economic analysis, with explicit references to published sources. In particular, he utilized the two special solutions of the restricted problem of three bodies found by Lagrange in 1772, an equilateral triangle

model and a straight line model, although he did not develop his discussion of the latter. Hotelling had used the linear model earlier, but had not revealed its origins. Neither had Edward Chamberlin (1962) in his critique of Hotelling; nor had Hoover. Following Pick, Dean showed that Hotelling's line model was a special case of the polygon model. As the sum of the interior angles of the locational polygon approached 0° or 180° , the polygon reduced to a line. The possibility of reducing a triangle to a line also is alluded to in Dzibek's book on planetary motions, and in Jeans' article (1954) on harmonic analysis.

Pick, born in Vienna, was ordinary professor of mathematics at the German University of Prague from 1892 to 1929. From 1929 to 1938, he served as rector of the University of Vienna. Thus, while Weber was at Prague from 1904 to 1907, they were colleagues, and in 1909 became collaborators on the economic theory of location.

Dean (p. 22) also found that by using the methods of the astronomers, he could conclude that 1) the parallelogram of forces theorem permitted the solution to the production location problem in spatial economic theory to be found at an interior site in the locational triangle, rather than at one of the vertices or along one of the sides, and 2) polygons with any number of vertices could be analyzed in the same way to solve an economic location problem.

Interestingly, using the parallelogram of forces to find the interior solution site as Dean did produces a diagram of a pyramid.

Dean's source for the conclusion that any polygon could be used for this purpose was not Launhardt or Weber/Pick, but Lynn Loomis, a Harvard graduate student in mathematics in 1938, who from 1941 to 1982 would be a faculty member in the Harvard Mathematics Department. Loomis was following the precepts of Birkhoff, Julian Coolidge and William Graustein, his professors in mathematics at Harvard. Dean's solution thus owed more to their topological bias than did those of Launhardt and Weber/Pick.

In 1953, Loomis published a book on harmonic analysis, a technique for analyzing periodic phenomena, which developed from

Pythagoras and his mathematical analysis of music, especially the problem of the vibrating string. This type of analysis has had a great significance in the search for the solution to the three body problem in the twentieth century, and in particular analyzes nodes. Poincaré and Birkhoff were especially notable contributors to this work. (Also see Yitzhak Katznelson, 1969; Jeans, 1954; Loomis, 1953; Marston Morse; Birkhoff; Weber; "Pick," 1959.)

Dean acknowledged the work of recently published location theorists, but stated that he had not altered his book to take into account their findings. It appears from his citations of economics faculty members under whom he studied, that agricultural economist John Black (1926) and economic historian Abbott Usher (1929, 1936) had the greatest impact on Dean's thought.

Dean's analysis clearly anticipated, and thus can be seen as the objective source of, the gravity or potential models in location theory developed beginning in the late 1940's by John Stewart (1948) and others.

(See Jeans, 1905; Scheffers, 1900; 1904; Walter Isard, 1960; George Zipf, 1946.)

Dean's analysis was in accord with the profession's experiment with analogies from physical science and mathematics in the 1920's and 1930's to find an objective way to analyze resource allocation choices. Thus, he was at the forefront of the experimentation in analytic technique in economics at that time.

The significance of Dean's approach can be gauged by reference to the work of fellow Harvard graduate student Paul Samuelson during this period. In 1947, Samuelson extended the use of the location theorist's mathematical and scientific sources to economics more generally, and made these methods central to conventional modern economic analysis, using several Birkhoff works. In the period prior to 1950, Samuelson is the only economist besides Dean to cite or otherwise explicitly refer to the work of the Harvard mathematicians and astronomers in celestial mechanics, and to use their results in his own work.

POST-COMMUNIST ECONOMIC TRANSFORMATION: HUNGARY VS. POLAND[†]

Institutional Legacies and the Economic, Social, and Political Environment for Transition in Hungary and Poland

By KEITH CRANE*

All the new governments of the former nations of the Warsaw Pact have committed themselves to creating market economies out of the shambles of the old centrally planned systems. On the eve of the revolutions that swept Eastern Europe in 1989, only the governments of Hungary and Poland had successfully altered some of the main features of the Soviet-type model. Because they had already eliminated a number of restrictions on private enterprise, introduced markets in some areas, and attempted to employ a few of the macro-economic instruments used in market economies to regulate economic activity, economic agents in Poland and Hungary have a familiarity with the institutions and functions of market economies lacking in the other countries.

This greater familiarity could lower the costs of transition to a full-fledged market economy. In the process of transition, old ways of doing things (activities or technologies) have to be discarded and new ones learned. New institutions, such as a commercial banking system and capital markets, have to be created. Because many economic agents in Hungary and Poland have already made mistakes that occur in learning how a market economy operates, fewer resources may now need to be expended in learning to adapt to the new market systems. On the other hand, the habits ingrained under the

old, reformed systems may make the transition more difficult. In particular, managers of state-owned enterprises may have become so accustomed to avoiding adapting to the continual perturbations in economic policy characteristic of the reformed Hungarian and Polish systems that they will focus their energies on lobbying to maintain subsidies, rather than seeking to make their operations more efficient. If this should be the case, countries such as Czechoslovakia or Bulgaria which are only now attempting to change their economies may incur lower costs in the transition, because economic agents, especially managers, will be less resistant to change.

This paper speculates on the advantages and disadvantages of past experience in economic reform by analyzing the nature and degree of exposure to markets in Hungary and Poland. Because economic policy changes of the enormity of those taking place in Eastern Europe are not introduced in a vacuum, the paper also discusses the economic, social, and political environment on the eve of the creation of the new, non-Communist governments. It concludes with a discussion of the implications of previous experience with reforms and of the environment in which reforms are introduced for the pace and sequencing of the transition to a market economy.

I. Legacies of Reform

Hungary. The Hungarian government adopted a package of changes in its economic system, the New Economic Mechanism (NEM), in 1968. This package of laws and institutional changes explicitly acknowledged many of the primary failures of the

[†]*Discussants:* Márton Tardos, Hungarian Academy of Sciences; Morris Bornstein, University of Michigan; John P. Hartt, Congressional Research Services.

*PlanEcon, Inc., 1111 14th Street, NW, Washington, D. C. 20005.

Soviet-type system; most notably, the absence of signals that would permit the efficient allocation of resources, the lack of incentives to efficiently utilize inputs, and the inability of the system to exploit potential gains from trade. The NEM attempted to remedy some of these problems by 1) giving enterprise managers' more independence and tying their remuneration to profits; 2) eliminating compulsory plan targets and allocating more resources through markets; 3) loosening price controls on a number of goods; and 4) introducing a single commercial exchange rate for enterprises.

After a period of retrenchment between 1973 and 1978 brought on by domestic economic problems, the deteriorating international economic environment, and political pressures, internally and from the other members of the Warsaw Pact, the government introduced a new round of policy measures in the 1980's designed to make the economy more market oriented (Jan Adam, 1987). The government created a two-tier banking system, restructured the price system so that it more faithfully followed relative prices on world markets, and liberalized foreign trade. The latter was accomplished through the relaxation or elimination of import controls, and by lowering barriers to entry for potential exporters. Beginning in 1982, the government also relaxed restrictions on the size of private enterprises and later permitted the creation of limited liability companies. In general, the climate for private business was much improved.

However, in contrast to the simultaneous introduction of a package of reforms in 1968, many of these measures were introduced separately and often for different goals; at times these measures worked at cross purposes. For example, in order to preserve current jobs for workers, the government expanded subsidies to prevent large, state-owned firms from going bankrupt. Because of pressures on the balance of payments, it frequently clamped on very strict import controls in the hope of narrowing Hungary's current account deficit (Janos Gacs, 1987).

In the year before the installation of the new government in 1990, the socialist government accelerated the pace of change.

The government passed laws permitting the conversion and establishment of joint stock companies, eliminated restrictions on private enterprise and foreign ownership, and liberalized most prices. It also passed laws on the privatization of state-owned property. In many ways the outgoing government took a more radical approach to economic change than has the current government.

Poland. Successive Polish governments made a series of attempts at economic reform (1956, 1972, and 1982). The last attempt has had more staying power than the previous two.

The 1982 reform began inauspiciously. The reform was implemented in January 1982, immediately after the imposition of martial law in December 1981. This reform was much more focused on the question of control than Hungary's NEM (*Kierunki reformy gospodarczej*, 1981). In Hungary, the ministry hired and fired the managing director. In Poland, enterprise councils, freely elected by the work force, were to decide on the directors, vote on corporate strategy, and dispose of the firm's assets. After the imposition of martial law, the government modified the original reform plans to give the ministries more control than the councils. Nonetheless, Polish enterprise managers were more beholden to workers and less beholden to ministries than their Hungarian counterparts (Jan Lipinski and Ursula Wojciechowska, 1987).

The reform shared the emphasis on trade liberalization, some price decontrol and more independence for enterprises of the Hungarian reform. It differed in the lesser role of markets in allocating consumer and intermediate goods, a consequence of misguided attempts to curb inflation through price controls. The Hungarian authorities wielded firmer control over the money supply than did the Polish. Consequently, inflation was lower and the government felt more confident that price liberalization would not degenerate into a more rapid rate of inflation. Thus, markets developed for a number of products in Hungary. The Polish government was much more willing to accept and monetize wage increases. Consequently, it clung to a series of ineffectual instruments, such as price controls,

to combat inflation. These policy instruments made it impossible for markets in intermediate goods and raw materials to develop, thereby preserving the old administrative system of allocating resources.

In contrast to Hungary, the last several months of Communist rule were disastrous in Poland. Although the Communist government established the equality of private and state enterprise and eliminated barriers to entry in most industries, it rapidly accelerated the rate of money creation, partly in an attempt to "buy" the 1989 elections. Inflation soared to year-on-year rates of over 700 percent.

II. The Economic, Social, and Political Environment

Despite previous experience with reforms, the new governments' policies were more directly affected by the economic, social, and political environment that existed upon the transfer of power.

Economic Environment. The new Hungarian government took power after a decade of very slow economic growth under the former Communist regime. The GNP grew at an average annual rate of 1.5 percent between 1980 and 1989. Although a substantial share of the population was able to supplement their incomes with private-sector employment and some entrepreneurs greatly increased their incomes, real wages fell 6.5 percent over this period. Furthermore, a large number of people fell below the official poverty line. To compound matters, inflation began to accelerate towards the end of the decade, reaching 17 percent in 1989.

The only dynamic sector of the economy in the 1980's was the private sector that accounted for all economic growth. Artisans doubled their share in net industrial output. The number of limited liability companies soared from 10,000 in 1988 to 23,000 in the first half of 1990. Newspaper accounts and anecdotal evidence suggest that the sector is considerably larger than indicated by official statistics, possibly as much as 30 percent of GNP (Janos Kornai, 1986). Payments for construction and repair services are usually made in cash and many of the self-

employed do not register with the government in order to avoid taxes.

Polish economic performance under the reform was little better than Hungary's. Productivity measures failed to improve, shortages persisted, and were generally more prevalent than in Hungary. Capital continued to be poorly allocated; rates of return on new investment were very low by international standards. By some measures, foreign trade performance improved. Exports became less concentrated in particular products and grew more rapidly than from other Warsaw Pact countries.

As in Hungary, the private sector in Poland exhibited extremely dynamic growth. Private sector employment in industry rose from 271,000 workers (5.2 percent of the industrial labor force) in 1980 to 717,400 workers (14.7 percent) in 1989. Private-sector output grew at commiserate rates.

Both countries faced severe balance-of-payments problems throughout the decade. After 1980, Poland failed to service its hard currency debts. Debts to commercial banks were eventually rescheduled; however, Poland refused to pay interest on debts to Western governments until Western sanctions were removed. Subsequently, interest on these debts was capitalized and loans rescheduled. However, these policies increased Polish debt from \$27 billion in 1981 to \$39.2 billion in 1989. Hungary remained solvent during this period, but increased its debt from \$10.216 billion in 1982 to \$20.605 billion in 1989. The government's policies were highly ineffectual in controlling the current account deficit.

Reforms failed to accelerate economic growth or improvements in factor productivity in either country in the 1980's. The socialist sector (cooperative and state-owned firms) performed very poorly. Among the many reasons for the failure of the reforms to accelerate economic growth, two stand out. 1) The state never created an effective mechanism to efficiently allocate capital (Gerhard Fink, 1982). It continued to make almost all investment decisions. Political pressures, not prospective rates of return, remained the determinants of capital allocation. 2) Enterprise managers were rewarded for their ability to respond to the

desires of their superiors in the supervising ministries, not on increasing profits, notwithstanding the statements embodied in the reform. Enterprises performed accordingly.

The Social Environment. Public opinion polls register greater concerns about crime in Poland. The decline in fear of the security apparatus and a general loosening of social strictures appears to have been followed by an increase in crime. Greater press freedom has probably resulted in more reportage. The combination led to a greater social concern about personal security.

In both Poland and Hungary, the elderly have been the most harmed by accelerating inflation, although government workers such as teachers and health professionals have also complained bitterly. The governments have not maintained the real incomes of the elderly through indexation.

Many people, particularly low-level bureaucrats and unskilled workers, became increasingly anxious about the possibility of unemployment. The combination of more rapid rates of inflation, the threat of unemployment, higher crime rates, and a general uncertain future has elicited two general reactions in both countries. A substantial share of people grasped the new opportunities, taking pleasure in the decline in regimentation. They have started their own businesses or tried to make the state-owned firms in which they work more competitive. However, another large group has felt more anxious and insecure in the emerging new system. For reasons of age or training, this group sees few prospects in the new system. Although they have not voted for the successors to the former communist parties, they have been fickle voters, seeking those parties which promise them some protection from the new economic uncertainties.

In both countries many indicators point to a decline in the quality of life over the past decade. Life expectancy for males has declined due to diet and higher levels of alcohol and cigarette consumption. Very high levels of air and water pollution have also contributed to lower life expectancies than in Western Europe.

Political Environment Although Poles and Hungarians shared many of the same social

problems and worries, the political environment on the eve of the transition was far different in the two countries. The opposition came to power in Poland after waging a long struggle against the government. It encompassed a strong labor movement and urban intellectuals, and had the general backing of almost the entire country as shown by Solidarity's sweep of the 1989 spring elections. This strong support made it possible for the opposition to take power despite its minority status in Poland's lower house, the Sejm. The opposition was therefore able to put together a list of candidates and a government that contained several well-known names with experience in public speaking and politics.

In contrast, the main Hungarian party, the Hungarian Democratic Forum, was midwived by Imre Pozsgay, one of the four leaders of the Hungarian Socialist Workers' Party. Founded in the fall of 1988, originally the Forum was a loose coalition of intellectuals and people who ascribed to the "populist" tradition in Hungarian intellectual and political thought that focused on the Hungarian peasant and the wellsprings of Hungarian nationality.

Hungary's second largest political party, the Free Democrats, was created by several of the dissidents that had badgered the Hungarian government since the 1970's. It ascribed to the more cosmopolitan, liberal tradition. It has formed an alliance with the Young Democrats, a more radical, outspoken group of young Hungarians that focuses on human rights issues and is strongly pro-market. Thus, in contrast to Poland where the opposition remained united against the communist regime, Hungary approached the changing environment with an incipient system of parties with clear policy differences.

III. Implications and Conclusions

Despite the shortcomings of the reforms, they appear to have given Poland and Hungary a leg up in the process of transformation, in part because some of the institutions necessary for a functioning, modern market economy had already been partially created, because of the greater familiarity of their inhabitants with market systems,

and probably most importantly, because past changes had led to the emergence of a large group of entrepreneurs with experience in opening and operating private businesses. In fact, roughly half of all Hungarians and a large percentage of Poles had worked in a private business or activity before the change of government. Thus, a tradition of working in and opening up private businesses has taken root.

The major negative experience with reform was that economic agents came to expect frequent modifications of government policies and regulations. However, governments of countries with more traditional socialist systems also frequently modified the rules so enterprise managers in all the countries had similar experiences in this regard. They learned that the most remunerative activity was to learn how to evade new regulations or lobby government officials so as to obtain favorable treatment.

The experiences of these countries with reform suggest that it is better to introduce a consistent package of reform measures at one time, rather than sequentially. In most cases all the policy measures are necessary to elicit the appropriate microeconomic responses. They also suggest that the governments need to continue to push institutional changes after the initial policy packages, so that government agencies begin to operate in a way conducive to the development of a market economy as well as the private sec-

tor. Finally, their experiences suggest that the quicker reforms are implemented, the better. Time and again, the Polish and Hungarian governments failed to introduce a measure because of "political" costs, only to introduce the measure later after economic conditions had further deteriorated, because they now felt they had no choice.

REFERENCES

- Adam, Jan, "The Hungarian Economic Reform of the 1980s," *Soviet Studies*, October 1987, 39.
- Fink, Gerhard, "Determinants of Sectoral Investment Allocation in Hungary," *Acta Oeconomica*, No. 3-4, 1982, 28, 375-88.
- Gacs, Janos, "Import Substitution and Investments in Hungary in the Period of Restrictions (1979-86)," in Andras Raba and Karl-Ernst Schenk, eds., *Investment System and Foreign Trade Implications in Hungary*, Stuttgart: Verlag, 1987.
- Kornai, Janos, "The Hungarian Reform Process: Visions, Hopes, and Reality," *Journal of Economic Literature*, December 1986, 24, 1687-737.
- Kierunki reformy gospodarczej*, Warsaw: Ksiazka i Wiedza, 1981.
- Lipinski, Jan and Wojciechowska, Urszula, *Proces wdrazania reformy gospodarczej*, Warsaw: Panstwowe Wydawnictwo Ekonomiczne, 1987.

Transformation Programs: Content and Sequencing

By FARID DHANJI*

Conventional wisdom will have it that Poland opted for a dramatic, compressed dash towards market capitalism—a “Big Bang” that in one stroke shed central planning and dismantled bureaucratic interference in large areas of economic decision making. Hungary, it is believed, has taken a more gradual, time-consuming, less forced approach towards the same end. These perceived contrasts in strategy (“shock therapy” vs. “gradualism”) are keenly debated in policy reform circles throughout the region, both with respect to necessity as well as to feasibility.

The conventional wisdom greatly simplifies the contrast between the two countries. Recent political developments in Hungary have generated a substantial acceleration of reform efforts. The Hungarian reform program now operates on as broad a front as the Polish program: abandonment of central controls, price liberalization, introduction of competition, privatization, financial and fiscal system modernization, development of social safety nets, legal, regulatory and institutional reform, and so on. It seeks a no less fundamental and far-reaching change.

The perception of gradualism probably has more to do with the past 20 years of Hungarian experimentation with piecemeal modifications to the economic mechanism (Janos Kornai, 1986) than to the present reality. That past experience certainly suggests that without genuine political commitment to a market economy, and particularly to a broad-based system of private property rights, systemic reforms in formerly socialist economies will not succeed.

Once these precepts are accepted, the specific agenda items for reform appear to take on the same character in all reforming

economies. The pace and sequencing of economic and institutional changes are then dictated by initial starting conditions that differ greatly between countries (Stanley Fischer and Alan Gelb, 1990).

Price liberalization and the dismantling of central coordination can be accomplished overnight; the establishment of new legal systems, privatization, the restructuring of enterprises, generating attitudinal and behavioral changes in populations, all perforce take much longer. As a result, it may be expected that the ambit for genuine choices on pace and sequencing on policy and institutional change will narrow over time and that the content of reforms in Hungary and Poland and other countries will increasingly resemble one another.

I. Macro Stabilization, Price, and Trade Liberalization

In essence, the January 1990 Big Bang in Poland was aimed at eliminating the near hyperinflationary conditions in 1989, occasioned by the prior complete breakdown of wage and credit discipline. (See David Lipton and Jeffrey Sachs, 1990, and Dariusz Rosati, 1990, for details.) The uniqueness of the program, however, and its dramatic implicit judgment on the shock therapy vs. gradualism debate, lay in the acknowledgement of the fact that under the new competitive rules of the game, significant unemployment would result from enterprise bankruptcies and redundancies. This was deemed acceptable, and necessary for economic restructuring.

The Big Bang program continued the substantial earlier progress achieved in jettisoning central coordination and liberalizing price setting. About 95 percent of producer and consumer prices are now freely set in Poland (Lipton-Sachs, p. 114). Trade was largely liberalized with very low tariffs set in the course of the year and the zloty was made convertible for most current account transactions.

*World Bank, 1818 H Street, NW, Washington, D.C. 20433. The views expressed in this paper are solely my own, and should not be attributed to the World Bank.

The results of the first few months of the program are now in, and there are some major surprises. On the positive side, first, there has been a practical elimination of shortages; second, the inflation rate has been substantially reduced from its preprogram peak of 50 percent per month, to 3–4 percent per month. On the negative side, measured output (that, unfortunately, neglects the contribution of the private economy) will likely drop 15–20 percent over the year as compared with an anticipated drop of about 5 percent. Unemployment of about 400,000 persons was expected; unemployment is now about 1 million, 7 percent of the labor force, and is expected to climb significantly higher. Real wages have fallen by about 30 percent (although the meaningfulness of this statistic is controversial, as it is measured against a base period when shortages prevailed). A major surprise has been a substantial export boom to convertible currency markets accompanied by a decline in imports; as a result, a trade surplus for the year is expected. A second major surprise has been a surge in enterprise profit margins. No bankruptcies have been observed.

Is this mostly “shock” and little “therapy”? As yet, no clear and plausible analysis has yet emerged—understandable in view of the novelty of the experience. Moreover, distinguishing between the effects of the necessary stabilization program, compounded now by the demise of the CMEA and the recent increase in oil prices, and the effects of the systemic structural reform measures, is virtually impossible empirically. While there is some anecdotal evidence that enterprises are responding creatively to the new market environment, for example, by abandoning unprofitable product lines, or by establishing new distribution outlets (Erika Jorgenson et al., 1990), it will be some time before it is known how pervasive and deep rooted the changes are, and how well enterprises are positioning themselves to deliver a strong supply response in the new market environment. The continuing price increases through the exercise of dominant positions in the face of declining demand must, however, give pause to any

over-sanguine view that a market culture is being systematically embedded in the state enterprise sector. At the same time, the rise of a new, flourishing private sector has been one of the most visible and promising aspects of the reforms.

Meanwhile, dissent is heard that the stabilization measures have comprehensively overshot their objectives (see Gregory Kolodko, 1990; Rosati) leading to urgent calls for domestic reflation. This position has its critics who point to the continued high levels of inflation, and to the considerable evidence that stabilization programs most often fail because they are not sustained. Political support for the program has also eroded. The election of Lech Walesa to the presidency may, in part, be attributed to popular fears of enterprise closures and further unemployment.

Hungary, too, has been undergoing a macro-stabilization effort accompanied by structural reforms. The stabilization has been aimed at reducing inflation, running at about 27 percent annually, and to correcting an external imbalance, essentially to avoid a rescheduling of Hungary's large external debt. On the latter count, the program has clearly been successful, as a balanced current account is expected this year. Output is expected to fall by about 6 percent in 1990, in good measure because of the supply shock engendered by reduced trade with the Soviet Union. The level of unemployment has increased to about 80,000 in November, about 1.5 percent of the labor force; it is expected to increase to almost 200,000 persons by the end of 1991, as industrial restructuring is accelerated. The fast expanding private sector is, however, rapidly absorbing labor.

The stubbornness of inflation in Hungary remains a puzzle. It appears less a result of loose monetary policy than of enterprises, hit by multiple shocks during the year (devaluation, reduction in subsidies, declines in ruble trade, administered price increases), responding by raising prices, and not reducing them in the face of slack demand. Lowering inflation appears to require measures to increase competition in the economy and to hasten the exit of large, chronically insol-

vent enterprises that act as a source of substantial demand pressure. Although a number of small bankruptcies have been seen, there have been no bankruptcies of large enterprises.

Price and trade liberalization have also continued apace. In 1988, the share of consumer goods whose prices were freely set was slightly over 50 percent; it is expected to exceed 90 percent in 1991. Similar progress is evidenced in the liberalization of producer goods prices where, again, about 90 percent of prices will be freely set in 1991. On the trade front, less than 10 percent of imports are expected to be subject to licensing by the end of 1991. The average tariff in Hungary is about 16–18 percent, and there is discussion of reducing this to half its current level.

In reviewing the stabilization, price, and trade liberalization approaches of Hungary and Poland, it seems clear that Hungary has fended off precipitous declines in output, employment, and wages more successfully than Poland. This would appear to have less to do with the approach Hungary has adopted, or the specific measures undertaken, as to the vastly differing initial conditions. Poland has been tackling a near hyperinflation in which the credibility of the old bureaucracy to manage a gradual transition was minimal, and where there were strong political demands to institute irreversible changes; Hungary has not had to deal with these conditions. The experience, therefore, does not appear to be conclusive to either side of the gradualism vs. shock therapy debates.

II. Privatization

The sequencing of privatization in the reform efforts in East and Central Europe has aroused considerable controversy. It has been argued that, without private owners interested in the returns to capital, there will be little incentive for the dominant state enterprise sectors to vigorously pursue profit maximization, cut costs, and otherwise restructure their overmanned and inefficient enterprises. In view of the reluctance of politicians to close down insolvent enter-

prises, continued subsidies and ultimately recourse to credit creation will be necessary, both undermining the stabilization efforts and putting the overall reform programs in jeopardy. Full scale privatization should logically, therefore, come early in the reform effort, and preferably precede other measures.

Whatever the merits of this view, it is clear that, both in Hungary and in Poland, privatization is proceeding slowly. Both countries had initially to contend with worker beliefs that they owned the enterprises they worked in, a belief fostered by the establishment of enterprise councils dominated by workers in the 1980's. In both countries, the state has had to reestablish its ownership claims.

Before the major political changes this year, both Hungary and Poland witnessed a spate of "spontaneous" privatizations, where, typically, managers would imaginatively exploit loopholes in legislation designed to transform enterprises into joint-stock companies, appropriate for themselves majority shareholdings in enterprises, or else lease enterprise assets to themselves for minuscule rents. The practice was quickly ended by the new Polish government. In Hungary, there remains considerably greater tolerance for the practice, although early "scandals" led to a strengthening of the supervisory powers of the State Privatization Agency.

Both countries are advancing in their plans to privatize enterprises. A number of sales of enterprises to foreigners has already taken place in Hungary; a further 20 enterprises were brought to the market in September, and the government intends to sell about 10 percent of state-owned assets annually over the next 3 years. In Poland, political debate in the Sejm over the Privatization Law took about 6 months to complete. Frustration over the slowness of the process surfaced when Walesa made this a major campaign issue. It appears that a new privatization modality will be introduced in which some fraction of enterprise assets, perhaps 30 percent, will be given free through a voucher scheme to the population, about 20 percent will be offered to

employees at a considerable discount, and the remaining proportion retained or sold by the state. Proposals have also surfaced for developing holding companies that will have supervisory managerial powers over enterprises in which state holdings are large. Whether this will lead to a substantial acceleration in the pace of privatization and, as importantly, to improvements in enterprise performance, remains to be seen.

III. Enterprise Restructuring

Socialist enterprises are traditionally highly intensive in the use of materials and energy, are considerably overmanned, and provide a number of social benefits to employees that are generally reserved for state provision in market economies. This has resulted in low levels of efficiency and competitiveness in settings where most enterprises enjoy monopolistic or near-dominant market positions.

In opening economies to market forces and in enforcing "hard budgets," it has generally been expected that many enterprises would fail, and the majority of enterprises would have to undergo considerable financial, technological, and manpower restructuring before they could survive in the new market environment. Different views are heard as to the urgency of this requirement, its position in the sequencing of policy actions, and especially as to who should undertake it. In Czechoslovakia, for instance, the Finance Minister has emphatically declared that the state bureaucracy has no comparative advantage in restructuring enterprises; this activity should be left to the new private owners.

Both Poland and Hungary have, however, allowed a role for the state in restructuring. The Hungarian experience is much longer, dating back to 1986, when the process of restructuring a number of enterprises engaged in foreign trade was initiated. Although progress since has not been rapid, an acceleration is likely to be seen now as the number of liquidation proceedings brought by creditors against insolvent enterprises has increased substantially; bankruptcy will now afford greater prospects

for restructuring. In addition, the government has announced plans to independently initiate restructuring in certain sectors.

In Poland, the pace has been slower. An Enterprise Restructuring Fund has been established, based on a modest tax on enterprises. A Development Bank has been created, that will take equity positions in enterprises and provide technical and financial assistance in restructuring to its clients. By and large, however, restructuring in the first year of the Big Bang has been left to enterprises themselves, and as noted earlier, it is not clear how much has been made of the opportunity.

Enhancing competition in both economies has generally been left to occur through trade liberalization. However, both countries now have Competition Laws on their books and have targeted specific sectors for the initial attempts at demonopolization. Action may be expected in 1991, and may afford considerable opportunities for simultaneously speeding the privatization effort.

Social Safety Nets. The building of social safety nets (unemployment compensation, job training and vacancy referral systems, social security and pensions) has long been considered an essential component of reform programs, not simply from humanitarian concerns, but also to ensure political support for the program, and to facilitate labor mobility. Poland has not yet instituted a comprehensive system of social support, although some social benefits were introduced in the first year of the program. Legislation is now being prepared to provide adequate income support for the unemployed, while providing incentives to return to work, combined with a program to enhance the employability of those unemployed. A draft law on other social benefits is also now being prepared.

While the Hungarian unemployment provisions have been somewhat more generous, the expected large increase in unemployment in the coming year will require a forceful effort to develop nonbudgetary mechanisms for financing safety net programs. A draft bill has been submitted to Parliament, embodying proposals for compulsory contributions to an unemployment

insurance fund from both employers and workers, and discussions are taking place about revamping the social security systems of pensions, disability benefits, family allowances, and health care contributions.

Other Reforms. Although a number of other reforms are under implementation or being initiated, two deserve special mention. First, in both countries, financial sectors are being reformed through the establishment of two-tier banking systems, with the central bank hiving off its commercial activities to commercial banks proper. In both instances, major challenges exist in capitalizing the banks, and writing off non-performing loans, as well as strengthening institutional development through improved regulatory and supervisory capacities, and upgrading skills and developing services.

Second, on the fiscal front, Hungary has already instituted a VAT as well as a corporation and personal income tax. Poland continues to use a turnover tax system, but has plans to institute a VAT. The preparatory work for this does, of course take a long time and the tax is unlikely to be introduced before 1993.

IV. Conclusions

A thumbnail sketch of the content of reform programs in Poland and Hungary, and the issues of sequencing that have emerged, can hardly do justice to the complexity of the reform effort or to the multifaceted dimensions of the issues involved. A brief evaluation of some of the emerging lessons might nonetheless be attempted.

First, it has been noted that reforms are a "seamless web" in which each element is intimately related to every other, sometimes as a necessary precondition, sometimes as a facilitating requirement, and sometimes as a fortifying circumstance. In both Poland and Hungary, reforms in dismantling the pervasive reach of the command system in price and trade liberalization, and in the movement to exchange rate convertibility, have proceeded rapidly. Privatization, enterprise restructuring, demonopolization, building social safety nets, and institutional strengthening have, for understandable reasons,

lagged behind. These reforms do, however, remain essential components of the strategy to elicit a strong and positive supply response, realign production structures, increase economic efficiency, and foster the reallocation of resources. Failure to move rapidly ahead in these areas will not only delay the hoped for production response, but also put great pressures on governments to relax financial discipline that, if indulged, can only result in undermining the reform efforts.

Second, as a corollary because reforms are all of a piece, countries do not enjoy the luxury of doing one thing at a time. In this sense, because some reforms by their very nature will take longer than others to implement, it is important to act on as many fronts at once, and sustain a critical mass of reforms in order to ensure success. Issues of sequencing may, perhaps, derive considerable theoretical appeal from assumptions of instantaneous application and effect, while ignoring the practical constraints of politics, administrative capacity, and the time taken to design programs and implement them. When these considerations are entered into the debate, most sequencing issues dissolve, and one is left with the more prosaic judgment that reform efforts require multiple simultaneous actions on a variety of fronts.

Third, despite initial appearances, Poland and Hungary in 1990 do not present a good study in contrasts. Their previous experience with reform efforts does, however, appear to demonstrate that political commitment to a vision of the market economy is necessary for reforms to advance, and that because of the strong interrelationships between elements of reform programs, it is preferable to advance rapidly rather than gradually. The lessons from both countries are less important to each other than to other countries that are beginning their reform efforts, especially perhaps, the Soviet Union.

Fourth, the high hopes for quick reforms and quick turnarounds in economic performance do not appear to be realistic. The experience of both countries suggests that reform is going to take a very long time, and reform efforts will need to be sustained over

many years. Moreover, the Polish experience suggests that the social and economic costs of reform now appear to be much greater than originally anticipated.

Finally, a point only briefly touched upon, political support for reform programs by the populations of reforming countries are an indispensable requirement for sustainability and eventual success. This obviously raises great challenges for the newly elected politicians in the emerging democratic settings, both to educate populations about the magnitude of the difficulties and the reasons for policies, while avoiding the temptations of populism. No plausible alternative to the agenda of reform actions outlined above has indeed emerged.

REFERENCES

- Fischer, Stanley and Gelb, Alan, "Issues in Socialist Economy Reform," mimeo.,
- World Bank, October 1990.
- Kolodko, Gregorz W., "Polish Hyperinflation and Stabilisation: 1989-90," Working Paper No. 10, Institute of Finance, Warsaw, 1990.
- Kornai, Janos, "The Hungarian Reform Process: Visions, Hopes, Reality," *Journal of Economic Literature*, March 1986, 24, 1687-737.
- Jorgenson, Erika A., Gelb, Alan and Singh, I. J., "The Behavior of Polish Firms under the Big Bank: Findings from a Field Trip," mimeo., World Bank, November 1990.
- Lipton, David and Sachs, Jeffrey, "Creating a Market Economy in Eastern Europe: The Case of Poland," *Brookings Papers on Economic Activity*, 1: 1980, 75-147.
- Rosati, Dariusz K., "The Sequencing of Reforms and Policy Measures in the Transition from Central Planning to Market: The Polish Experience," mimeo., OECD, November 1990.

Foreign Economic Liberalization in Hungary and Poland

By PAUL MARER*

Economic liberalization broadly defined means reducing the role of the authorities in microeconomic decisions, increasing reliance on the price mechanism rather than controls, establishing or strengthening market institutions, and integrating the country into the world economy. This paper deals with the last issue.

There is a close interconnectedness between liberalizing a country's foreign economic relations and changes in its domestic economy. Foreign economic relations that must be liberalized during the systemic transformation encompass relations among the countries of the Council for Mutual Economic Assistance (CMEA),¹ imports from convertible-currency areas, the convertibility of the currency, and foreign direct investment. I will focus on the first three issues.

I. Revamping the CMEA

The institutional mechanism of the CMEA that was in effect from 1950 until the end of 1990 was not compatible with the decentralized, market-driven economy toward which Hungary and Poland wish to progress.

Most trade decisions with CMEA partners were made by governments, bilaterally. Because domestic costs and prices were not market determined, and because the prices and the quantities of individual exports and imports were linked, it was not possible routinely to determine whether a transaction was in accordance with comparative advantage, especially for manufactures. Therefore, even in reformed economies such

as Hungary's, the state had to insure through taxes and subsidies that each transaction was profitable, up to a point, to each firm producing exports or using imports. There was no rationale to reward firms for improving the composition, quality, modernity, or after-sales service of exports, since it was very difficult to obtain from the trade partner's government a correspondingly better price or some other quid pro quo. This ossified the structure of intra-CMEA trade and weakened the abilities of firms to compete on the world stage.

Trade was conducted and settled in transferable rubles, a unit of account that was neither convertible nor transferable. Trade had to be bilaterally balanced. There was no point in generating a surplus whose value was uncertain, since it could be settled only by future shipments by the debtor. If the debtor country had goods it was willing to ship and the creditor wanted to buy, there would have been no surplus in the first place.

Yet, the Central and East European countries did benefit by being large net exporters of overpriced manufactures, and large net importers of underpriced energy and (some) raw materials in their trade with the Soviets. But these easy gains helped to preserve East Europe's excessively energy- and material-intensive structure of production and dependence on the Soviet Union as a supplier and as a market. The USSR was willing to incur some opportunity costs to trade with Eastern Europe in order to tie its trade partners into the Soviet-led Warsaw Pact political-military alliance.

Two aspects of intra-CMEA arrangements are separable. One is the system of pricing and settlement, that can be moved to current world market prices and settled in dollars without having to transform either partner's economic mechanism. The other is the system of foreign trade decision making. This is much more difficult to change be-

*Professor, International Business, Indiana University, Bloomington, IN 47405.

¹The CMEA was founded in 1949 by the USSR, Bulgaria, Czechoslovakia, Hungary, Poland, and Romania. East Germany joined in 1950; Mongolia, Cuba, and Vietnam later.

cause who makes foreign trade decisions and on what basis depends in large part on a country's domestic economic system. As long as, say, the USSR remains centrally planned, can Hungary or Poland set free their own enterprises to swim or sink in trading with the Soviets?

In the late 1980's, these issues were debated. It was clear that rapid transition to a new pricing, settlement, and trading system would create major shocks in Hungary, Poland, and the other economies of Central and Eastern Europe. At the same time, the changes would also help teach their firms how to swim, that is, to compete on the world market. Different CMEA countries, and experts in those countries, had different views on the tradeoff between the short-term costs and long-term gains of transition to alternative types of new arrangements (as detailed in Andras Kovcs, 1991).

In 1989, Hungary became the first country that, after arduous internal debates, officially recommended such a change, to be applied first in trade with the Soviets. But before Hungary could negotiate the full details of a new arrangement, in June 1990, Gorbachev announced that, beginning in 1991, Soviet trade with the CMEA countries will be valued at current world prices and settled in dollars. As the political rationale of keeping the Central and East European countries tied to the USSR has greatly diminished, and as the Soviet economy came to face immense difficulties, the main aim of the Soviet Union became to improve its terms of trade and earn more convertible currency.

A system of world market pricing and dollar settlement is equally compatible with profit-driven trade, with deals decided by firms alone, or with continued government involvement in enterprise decisions, without granting to firms the right to convert their earnings freely into commodities or hard currency. Soviet ideas seem to be closer to this second option. But as the Soviet Union is spiraling into ever-greater economic and political dislocations, individual republics, regions within the republics, and enterprises themselves are increasingly making their own barter foreign trade deals. These deals

are sometimes with counterpart agencies in the other countries, but are most often with individual exporters and traders.

During 1990, Poland's and Hungary's levels of intra-CMEA trade declined between 20 and 30 percent, not only with the Soviet Union but also with the other Central and East European countries. The largest decline was registered in trade with the former East Germany, where sales by West German firms supplanted much of what was formerly imported from CMEA partners. In Hungary and Poland, the authorities have imposed on firms restrictions on exporting to the CMEA, to prevent the accumulation of dubious credits.

The future of intra-CMEA trade is highly uncertain. Difficulties in the Soviet Union and in the other CMEA countries, together with the switch to dollar pricing and settlement, are sure to cause further large declines in trade volumes in 1991 and perhaps beyond. The system of trade is also uncertain; most likely, the arrangements will remain *ad hoc* and varied.

Apart from the decline in Hungary's and Poland's intra-CMEA terms of trade (with rough estimates of the loss running in the neighborhood of \$1 billion for each country), an even more serious adverse impact is the domestic multiplier effects of the declines in trade volumes. Although both countries will continue to reorient their trade to the West, the collapse of intra-CMEA trade and the change in the terms of trade will have large negative impacts on their economies in the short and medium run. But, if the new arrangements and pressures force enterprises to improve efficiency, change their production structure, and find new markets, then the collapse of the old CMEA can help set the stage for sustained economic recovery and integration into the world market.

II. Liberalizing Imports from the West

Until the late 1980's, Hungarian, Polish, and other Central and East European firms were almost totally protected from foreign competition. Protection was exercised not by quotas and tariffs, but through a complex and informal system of import rationing.

The logic of the system meant that all these economies pursued extreme versions of import-substitution industrialization. But, because building and operating the new capacities that were constructed to replace imports from the West required ever larger volumes of investment goods as well as basic and intermediate imports from the West, their foreign trade participation ratios and external dependence on the West increased, even though their policies intended to reduce dependence.

Owing to formal and informal domestic price controls, scarce imports from the West were not priced higher than domestic products or those imported from the CMEA countries. Thus, one cannot get a sense about the degree of protection in these countries by comparing prices on the domestic and foreign markets, as we do in market economies.

Another difference between Hungary and Poland on the one hand, and market economies on the other, is that in the former countries, a number of important "liberalization" steps have to be taken before import liberalization can begin in the sense that we use the term (i.e., allowing imports to compete with domestic production). One such step is ending the complete insulation of production from foreign trade. That is, importers should pay the actual costs and exporters receive the actual proceeds of their foreign transactions, and not the fixed domestic price, as is the practice under traditional central planning. In Hungary, this step was taken in 1968; in Poland, partly in 1972 and partly in 1982.

Another important "preliberalization" step is ending the monopoly of foreign trade, that is, the right of a single state-owned foreign trade company to conduct all export and import business in a given commodity, which is another feature of central planning. Such a system is usually dismantled gradually, first by giving foreign trading rights to a few, and then to a larger number of producing firms, and then by permitting first a few and then a larger number of foreign trade enterprises to compete against each other. In Hungary, the first halting steps were taken in 1968 when a handful of

industrial enterprises received exporting rights. In 1979, limited competition between foreign trading firms was introduced. The opportunity to engage in foreign trade as a right, not as a privilege, was granted to firms only recently. In Poland, this process began in the 1970's and proceeded slowly until December 1988, when a new law was passed, stating that every type of business is permitted that is not explicitly forbidden. In January 1990, virtually all remaining elements of the state monopoly of foreign trade were eliminated.

Import liberalization in a traditional sense began in Hungary in 1989. A 3-year program was adopted, at the end of which 80 percent of actual imports would be liberalized. To be sure, since a substantial segment of imports (in Hungary, Poland, and all the other Central and East European countries) are comprised of raw materials and parts that have no domestic substitutes, the share of domestic production that will be exposed to import competition will remain considerably less than 80 percent. Nevertheless, Hungary's move, taken at the time of severe balance-of-payments pressures, was pioneering.

Hungary's tariff levels and structures are close to the world average. The implicit quantitative restrictions (QRs) that characterized the old system were not converted into equivalent tariffs, even though an argument can be made that the authorities should have done so, temporarily, to ease the shock of long-protected enterprises suddenly being exposed to foreign competition. Nor were any extra tariffs or QRs introduced as instruments of industrial or trade policies, as they often are in the developing countries. Nevertheless, during 1989-90, imports did not surge because liberalization coincided with a soft domestic economy, a sharp decline in exports to the CMEA, and tight monetary policy.

In Poland, imports were liberalized in one fell swoop in January 1990. At the same time, a new import tariff was introduced, with an average *ad valorem* rate of 12 percent; on manufactures, 21 percent, with a wide dispersion of the rates. In addition to customs duties, imports also bear a hefty

turnover tax, replaced as of January 1, 1991, by a uniform value-added tax. Initially, some protection was also provided by the exchange rate.

The new customs law recognized that the impact of sudden competitive pressures may have to be mitigated in some sectors. Therefore, provisions were made for the easy temporary imposition of QRs, various surcharges, or simply higher tariffs. During 1990, such interventions did occur, for example, on imported cars and consumer electronics.

Fears that rapid import liberalization and the introduction of convertibility would lead to massive trade deficits did not materialize. On the contrary, Poland's convertible-currency trade moved from a small surplus of about \$100 million in 1989 to a whopping estimated surplus of \$3.5 billion in 1990. This was attributable to similar factors as those mentioned for Hungary, reinforced by extremely tight fiscal policies. During the first half of 1990, the accumulated budget surplus reached about 20 percent of total government expenditures, about 6 percent of the GDP. In turn, the surplus can be traced to 1) high profits of the enterprise sector, the result mainly of price liberalization under a highly monopolistic market structure; and 2) wage increases in the budget-funded sectors much below the rate of inflation.

III. Convertibility

Convertibility enhances efficiency because it helps to eliminate price distortions if domestic pricing and foreign trade policies are liberal enough to permit the importation of the international price structure. Under these conditions, most benefits of improved efficiency can be obtained if convertibility is assured for the domestic enterprise sector for current-account transactions.

With the substantial recent liberalization of imports (which means that enterprises that have the domestic-currency cover can purchase foreign currency freely to import liberalized goods), Hungary has taken a big step toward *de facto* convertibility of the kind mentioned. In addition, the National Bank of Hungary guarantees the convert-

ibility of profits that a foreign investor wants to repatriate, and also that of foreign capital invested, with certain restrictions. In 1988, generous foreign-currency travel allowances were provided to all residents. But, there was such a surge of travel (shopping tourism, mostly to neighboring Austria) that severe restrictions had to be reimposed to protect Hungary's precarious balance of payments. Thus, for the man on the street, the forint is not a convertible currency, even though it has many of the attributes of convertibility for resident enterprises and foreign investors.

The exchange rate of the forint is set by the National Bank of Hungary and is pegged to a currency basket. Concerned about inflation, that during the last several years has accelerated from single digit to about 40 percent in 1990, the Bank has maintained between 1987 and 1991 a real effective exchange rate (nominal effective exchange rate adjusted for relative movements in domestic prices and prices in the main trade-partner countries) that is considerably higher (fewer forints per dollar) than the marginal rate or the rate that the IMF would like Hungary to maintain. Export subsidies are thus still required.

In April 1989, Poland removed many restrictions on foreign-currency operations and introduced limited foreign exchange auctions. But the really important changes were introduced on January 1, 1990: a nearly 50 percent devaluation, from 6,500 zloty to the dollar to 9,500 to the dollar, pegging the zloty to the dollar, and making the currency internally fully convertible for all current-account transactions, without distinction between the enterprise and the household sectors. Western governments put together a \$1 billion stabilization fund (to which the United States contributed \$200 million) to support convertibility at the pegged exchange rate.

During January–May 1990, consumer prices rose 160 percent. This meant a substantial real revaluation of the zloty. But because 1), the initial devaluation allowed for *some* of the increase in prices; 2), highly restrictive monetary and fiscal policies, together with continued high inflation, caused large increased demand for zloty cash bal-

ances by enterprises as well as households; and 3), the large real depreciation of the dollar against other Western currencies meant that the zloty was de facto also depreciated, the exchange rate had remained stable throughout 1990, without the authorities having to impose foreign-exchange controls, and without a penny of the \$1 billion stabilization fund having to be touched.

To be sure, the fact that in 1990, Poland paid only about one-fifth of the interest due on its approximately \$44 billion foreign debt, and thereby "saved" having to buy foreign exchange equivalent of about 5 percent of GDP in foreign currency, had also contributed to the strength of the zloty.²

²Since 1981, Poland has not paid interest on debts owed to Western governments, which now stands at close to \$30 billion. Practically all of the \$15 billion increase in Poland's debt between 1981 and 1990 was due to this unpaid interest being capitalized. Early in 1990, the Polish government, assuming (incorrectly, as it turned out) that the country would run a trade deficit in 1990, made a unilateral decision to pay no principal or interest on its medium- and long-term commercial bank credits that were not rescheduled. However, it is servicing fully and promptly its rescheduled debts and trade credits. Poland's decision vis-à-vis the commercial banks has impaired its creditworthiness, discouraged Western banks from setting up branches in Poland, and slowed considerably the flow of Western direct investment.

The difference between Hungary, that serviced its large foreign debt fully, and Poland, that did not, is important to keep in mind in any economic comparison that one makes between the two countries.

IV. Conclusions

Poland's runaway inflation required a shock treatment to bring it under control, whereas Hungary's macroeconomic imbalance was not so great as to require such a treatment. Foreign economic liberalization was a key aspect of the package of stabilization measures Poland introduced in 1990; the very same measures are also important building blocks for the systemic transformation. In these areas, as well as on arrangements replacing the old CMEA mechanism, Poland and Hungary are moving on a rather similar path.

REFERENCE

- Koves, Andras, "Transforming Commercial Relations with the CMEA: The Case of Hungary," in his and Paul Marer, eds., *Foreign Economic Liberalization: Transformations in Socialist and Market Economies*, Boulder: Westview, 1991.

Derivation of "Rational" Economic Behavior from Hyperbolic Discount Curves

By GEORGE AINSLIE*

Recent research has discovered frequent anomalies in the utilitarian reasoning of the normal human adult (Amos Tversky and Daniel Kahneman, 1981; Richard Thaler 1987). One of these seems especially inimical to a rational market economy: the finding that peoples' preference for a smaller good vs. a greater but more delayed good often changes as a function of the time the choice is made, even though the difference in delay stays constant. For instance, a majority of adults report that they would rather have \$50 immediately than \$100 in 2 years, but almost no one prefers \$50 in 4 years over \$100 in 6 years, even though this is the same choice seen at 4 years' greater distance (see myself and V. Haendel, 1983, for a systematic study).

Such change of preference as a function only of elapsing time is not an isolated finding, but has been observed in undergraduates choosing between longer or shorter periods of access to a video game or relief from noxious noise; in women deciding whether or not to have anaesthesia for childbirth; in substance-abuse patients choosing between different amounts of real money; and even in animals choosing between two amounts of food at different delays (see my forthcoming study, ch. 3).

According to conventional utility theory, the value of delayed goods is discounted in an exponential curve; the curves from two alternative amounts of the same good available at different times should never cross in the absence of new information. How-

ever, when behavioral psychologists have conducted parametric studies of choice, they have found a radically different discount function that has become known as Herrnstein's matching law. Various experimental designs have given the curve a number of specific forms, but all are hyperbolic—more bowed than an exponential curve, so that preference for goods of different sizes at different delays will indeed change as a function of time. The best version of this discount curve for a single good is probably J. E. Mazur's (1987):

$$(1) \quad V = A / (\zeta + \Gamma(T - t))$$

where V is the good's value (its ability to compete with alternatives), A is its amount, T is the time at which each good is available, and t the time of the behavior that obtains it (so that $T - t$ is its delay from the moment of choice), ζ is an empirical constant that determines value at zero delay, and Γ is an empirical constant that modifies the steepness of the delay gradient. In the limited data available, neither constant seems to range far from 1.0.

If formula (1) is used to compare two goods separated by three units of time, the later good objectively twice as large as the earlier, the indifference point (where $V_{\text{later}} = V_{\text{earlier}}$) occurs when the earlier good is available two units of time from the moment of choice. At a delay of five units, the larger good is preferred by $2/(1+5+3) = 2/9$ to $1/(1+5) = 1/6$, while at one unit the larger good loses by $2/5$ to $1/2$. The smaller good is temporarily preferred to the larger when the delay is short.

Insofar as choice is governed by the matching law, a tendency to form temporary preferences will present a major obstacle to rational planning: Any plan requiring a pro-

*Veterans Administration Medical Center, Coatesville, PA 19320 and Jefferson Medical College. Suggestions by Elmer Schaefer, Margo Schaefer, George Loewenstein, and Drazen Prelec are gratefully acknowledged.

longed course of action will fail unless the person can arrange consistent motivation for or binding commitment to it. Looking at the long view, he may want to be generally thin, brave, and prudent, but to accomplish this, he will have to overcome strong desires for food, escape, and financial abandon in the immediate future. Ulysses and the Sirens will not be a remote fantasy but a central problem of life.

On the other hand, financial markets display a different pattern of human choice making. Participants behave as though they discount future goods at single digit, exponential rates, and they do so without having bound themselves to any obvious mast. But then, which is the "true" discount function—the matching law or the exponential rate available from the bank?

The evidence suggests some kind of mixture. Depending on the circumstances, ordinary people choose annual discount rates in the thousands (see my paper with Haendel) or hundreds (George Loewenstein and Thaler, 1989) of percent as well as the bank rate, and can be seen adopting committing devices that resemble the psychoanalysts' "defense" or "coping" mechanisms (my forthcoming study, ch. 5), as well as engaging in straightforward rational planning. Furthermore, a careful look at the implications of temporary preference formation suggests mechanisms by which a person who evaluates goods strictly according to the matching law can be expected to arrive at the banker's shallow exponential discount curves for at least some of his transactions, but imperfectly, as a result of varying skill and effort. These will be my topic.

I. Stability via Detection of Intertemporal Prisoner's Dilemmas

This mechanism is the personal rule, that can be derived from the matching law as follows: If an individual must make a series of choices between goods of amount A_i and later, larger goods of amount A'_i (i.e., all $A'_i > A_i$ and all $T'_i > T_i$), each choice will be described simply by the matching law evaluation of the two alternatives involved in that particular choice unless the choices are linked. If, however, the whole series of

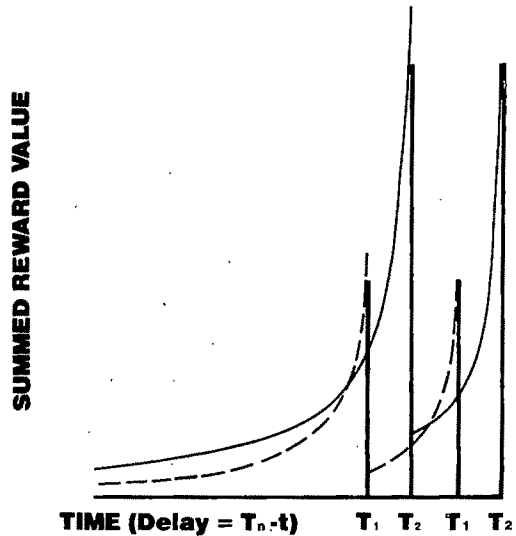


FIGURE 1

choices must be made all at once in the same direction, then the choice will be governed by the summed values of the goods on each side. Looking at ratios of summed values, each derived from formula (1), the crucial time at which preference between the two whole series of goods changes will be represented by the t when the value V' of the series of larger goods equals the value V of the series of smaller ones, called t_{indif} :

$$\frac{V}{V'} = \frac{\sum A_i / (\zeta + \Gamma(T_i - t_{\text{indif}}))}{\sum A'_i / (\zeta + \Gamma(T'_i - t_{\text{indif}}))} = 1$$

If the choice is made before t_{indif} , it will favor the series of larger, later goods, and if it is made after t_{indif} it will favor the series of smaller, earlier ones.

This would be a trivial application of the matching law to the case of serial choices except for an important phenomenon: t_{indif} between the series of larger (primed) goods and the series of smaller (nonprimed) ones will move closer to the moment when the first smaller good is available as the series is made longer (see my 1975 article). The period of temporary preferences for the smaller good will be reduced or eliminated. Figure 1 shows this for the simplest case, two pairs of goods with both $A_i = A$ and both $A'_i = A' = 2A$. The curves show their

values at all times t before they are due. The period of temporary dominance by the smaller, earlier good is shorter when the curves from another pair are added (just before the left-hand pair) than when only one pair remains available (just before the right-hand pair). With appropriate mathematical transformations, this finding holds equally for streams of continuous reward as for discrete, momentary goods.

The practical effect of choosing a whole series of goods at once is thus to increase the individual's tendency to choose the larger goods. That is, a given small, early good can be available more imminently without his forming a temporary preference for it. This predicted phenomenon is obviously relevant to the problem of intertemporal consistency.

But how can a person arrange to choose whole series of goods at once? In fact, he does not have to physically commit himself. The values of the alternative series of goods cannot depend on whether he will actually get them, an event that has not yet occurred, but only on his expectation of getting them. Assuming he is familiar with the outcomes of his possible choices, the main element of uncertainty will be what he himself will actually choose. In situations where temporary preferences are likely, he is apt to be genuinely ignorant of what his own future choices will be. His best information is his knowledge of his past behavior under similar circumstances, with the most recent examples probably being the most informative. Furthermore, if he has chosen the smaller good often enough that he knows self-control will be an issue, but not so often as to give up hope that he may choose the larger goods, his current choice is likely to be what swings his expectation of future goods one way or the other. If he makes an impulsive choice, he will have little reason to believe he will not go on doing so, and if he controls his impulse, he has evidence that he may go on doing that. The same logic is the basis for what is called a "self-enforcing contract" between individuals (B. Klein and K. B. Leffler, 1981).

According to this logic, amplification of impulse control can be expected to occur to

some extent whenever a person perceives a series of confrontations with temptations as similar to each other. He will not necessarily notice the process itself, or develop any way of describing it. He may develop an extensive practical understanding of it by trial and error, but have only tangential theories about how it works. However, insofar as he has become aware of this phenomenon, he will be able to induce it where it has not occurred spontaneously, by arbitrarily defining a category of gratification-delaying behaviors that will thereafter prevail or not as a set (see my 1975 article).

Such personal rules can be seen as a solution to a bargaining problem. The temporary preference phenomenon creates a relationship among an individual's successive motivational states that can be described as limited warfare (T. C. Schelling, 1960): Successive motivational states have some interests in common, and others that are peculiar to them. The interests in common are identical with the person's long-range interest. The peculiar ones are short-range interests in whatever goods happen to be imminently available. At any given time, the alcoholic wants to drink less in the aggregate (he does not want to be an alcoholic), but he may want to drink a great deal currently. His long-range interest, common to all his successive motivational states, is to be generally sober; this interest is challenged, and often overwhelmed, by a succession of short-range interests in getting drunk just once.

To explore an everyday example: Say that a person at midnight faces the choice of staying up for two more hours and having fun before he finally gives in to fatigue, but feeling tired at work the next day, vs. giving up his present fun and expecting to feel rested at work. He values the imminent fun at 60 units per hour, and expects to lose 60 units per hour of comfort from when he gets up at 7 A.M. until he leaves work at 5 P.M. At midnight, the value of staying up will be

$$V_{up} = \sum_{i=0.5 \rightarrow 1.5} 60/(1+i) = 64$$

and the differential value of feeling rested at work will be

$$V_{\text{bed}} = \sum_{i=7.5 \rightarrow 16.5} 60/(1+i) = 49.$$

Given only this one choice, he will stay up and suffer the next day. However, if he faces this choice nightly, he may perceive his current choice as a precedent for future nights as well. Assuming he believes that he will go to bed on time on subsequent nights if he does tonight, and not otherwise, the values of his alternatives are

$$\begin{aligned} V_{\text{up}} = & \left(\sum_{i=0.5 \rightarrow 1.5} 60/(1+i) \right) \\ & + \left(\sum_{i=24.5 \rightarrow 25.5} 60/(1+i) \right) \\ & + \cdots \left(\sum_{i=216.5 \rightarrow 217.5} 60/(1+i) \right) = 78 \end{aligned}$$

for staying up on the next 10 nights, vs. (by similar calculation) 105 for going to bed early on the next 10 nights. He will go to bed, *if* he expects that he will thereby be motivated to follow suit on the subsequent nights.

Considering separately the present values of the alternatives in his first choice (64 vs. 49), and the present values of two subsequent series of 9 choices all in one direction (14 for always staying up vs. 56 for always going to bed), his incentives create a prisoner's dilemma between his own present and future motivational states (Table 1), one that he will face on a nightly basis. From his present point of view, going to bed on time both today and in the future is worth 105, and staying up both today and in the future is worth 78. However, if he can stay up today and still expect to go to bed on time in the future, that is worth 120. Conversely, if he goes to bed today but fails to also go to bed in the future, that is worth only 63. The latter two outcomes respectively represent his successfully finding a loophole that excepts the present case from the string of precedents, and falsely hoping

TABLE 1

Now	Future	
	Stay Up	To Bed
Stay up	64 + 14 = 78	64 + 56 = 120
To bed	49 + 14 = 63	49 + 56 = 105

that his current cooperation would give himself sufficient incentive to follow suit in subsequent choices.

The person's best move at present will thus depend on how he forecasts his future perceptions. Insofar as he sees his current choice as a precedent and not an isolated instance, he will face the incentives of a repeated prisoner's dilemma.

II. Factors Stabilizing a Person's Valuation of Money

The incentives to cooperate in an internal prisoner's dilemma will take us part of the way from the sharp fluctuations of matching law discounting to the prudence of rational investment. It is likely that valuations of financial transactions start out the same way as the valuations of the visceral goods like alcohol and sleep that have just been mentioned. Assume for instance that a person likes to skimobile in the winter and sail in the summer, that he is just willing to pay what each vehicle costs (\$1000 for a used model on which there is little annual depreciation) at the beginning of the season, and that he can sell each back to a store at the end of the season for 25 cents on the dollar. Then every 6 months, he will face the choice of getting \$250 immediately vs. saving \$1000 in 9 months, alternatives that formula (1) evaluates at \$250 and \$100, respectively, if $\zeta = 1$ month. If he actually sees this as an isolated instance, the matching law predicts that he will sell.

However, if he expects to face a similar choice twice a year for roughly the next 20 years, the incentives he faces are the sum of 40 \$250s, one immediate, the others discounted, vs. 40 \$1000s, each 9 months after the corresponding \$250. These contingencies, too, represent a prisoner's dilemma

TABLE 2

Now	Future	
	Sell	Hold
Sell	$250 + 167 = 417$	$250 + 489 = 739$
Hold	$100 + 167 = 267$	$100 + 489 = 589$

(Table 2). He will hold his equipment for next season, *if* he believes he must do so this time in order to continue doing so (i.e., if the upper right cell is not a credible outcome), *and* that he will continue to do so if he does so this time (i.e., if he does not see a great risk that the lower left outcome will occur).

Seeing a transaction as a member of a larger category somewhat dampens the fluctuations in spontaneous value predicted by the matching law. However, even the person who does this in the above example will be indifferent between selling and holding at a selling price of \$354, for a good that will be worth \$1000 to him in just 9 months. While people may sometimes buy and sell according to discount rates of such magnitude (the effective rates in Hausman's study of air conditioner purchases reached 89 percent; see Loewenstein and Thaler), it is still much higher than would be called either normal or prudent. We must appeal to three additional factors to produce the stability with which economists are familiar.

First, cash pricing makes a wide variety of transactions conspicuously comparable, and hence invites an encompassing personal rule about the value of money generally. Just as the person in the foregoing example achieved more constancy than he otherwise would by seeing his choices about the skimobiles as related to his choices about the sailboats, so he will become more constant still if he sees each of his financial transactions as a precedent for all others. That is, if he sees what he spends for food, clothes, movie tickets, toys, postage stamps, etc., as examples of wasting or not wasting money, he will add thousands of examples to his interdependent set of choices, each flattening his effective discount curve by a greater or lesser amount.

The ease of quantitatively evaluating and comparing all financial transactions lets the value of purchasable goods fluctuate much less over time than, say, the value of an angry outburst, or of a night's sleep. Accordingly, it is common to see someone swayed considerably more by today's emotional comfort than by next year's, but also common to see him behave as if today's wealth were worth only a tiny fraction more than next year's. However, an encompassing rule creates as a side effect the familiar sight of people pinching pennies lest they set a wasteful precedent. The stability that it brings about is apt to be insensitive to cost effectiveness.

Second, financial transactions tend to become rivalrous activities. This adds an additional stake to the intrinsic consumption value of the goods involved in these transactions. Unless the night owl robs his sleep so much that he snoozes at work, his discomfort is unlikely to become the basis for invidious distinctions between himself and his fellow citizens. However, in buying and selling, he is not choosing simply in parallel with his neighbors, but in competition with them. If some of them are prudent enough to buy his skimobile every spring for \$500 and sell it back to him every fall for \$1000, they will soon wind up richer than he, and rewards in power over him, not to mention the sweetness of victory, will be added to the goods that originally seemed to be at stake. Of course, rivalry may also make people rash; but if the relevant choices are perceived in series of precedents as above, they will further enlarge the motivation for deferral of gratification. It may be the need to defend at least rough comparability with one's peers that makes the credit card rate of 20 percent the indifference point for consumer interest, instead of 200. Conversely, it may be the relative invisibility of Japanese growth to American consumers that lets the latter continue to indulge in a 20 percent rate rather than setting a still lower one, despite the consequent transfer of wealth overseas.

Of course, a society often makes nonfinancial activities a basis of competition as well. Where people gain an advantage by

staying hungry to attain stylish slimness, or by cultivating sexual indifference to increase their bargaining power with partners who are more readily aroused, the personal rules governing these activities gain power from these additional stakes just as the rules governing the value of money do, and can sometimes motivate heroic acts of abstinence. However, just as cash pricing labels the largest number of a person's choices as comparable, so it engages the largest number of people in social competition.

Finally, a person can set up his personal rules so that investment decisions are not weighed against his strongest temptations. As H. M. Shefrin and Thaler (1988) have recently pointed out, people assign their wealth to different "mental accounts" such as current income, current assets, and future income. These accounts seem to represent personal rules for how readily the money they govern may be used to satisfy immediate wants (see my forthcoming study, ch. 7). Skillful designation of income as "investment" vs. "spending" money (or even less protected "pin" money) may permit a person to accept bank discount rates on some of his money, while still appeasing temptations that might overwhelm a monolithic rule. Thus an individual may evaluate goods in compartments, requiring investment rationality in one, abandoning himself to the matching law in another, and probably following intermediate rules in still others.

However, temptation can defeat this strategy not only by sheer intensity or immediacy (as some drugs notoriously do), but by motivating the search for loopholes in the rules that assign money to one account or another. In Shaw's *Pygmalion*, the ne'er-do-well Alfred P. Doolittle asks Professor Higgins for five pounds to spend on "one good spree for myself and the missus." The professor is charmed by Doolittle's hedonism and offers him ten, but Doolittle refuses. "Ten pounds is a lot of money: it makes a man feel prudent like; and then goodbye to happiness." Most rationalizations are less costly, but may still erode a person's will to maintain an account governed by conventional prudence, leading

perhaps to the compulsive spending syndromes that are so often observed.

Taken together, the mechanisms just discussed may account for why priceable goods are not usually evaluated simply according to the matching law, even though that law may underlie all valuations. I have argued that goods are assigned value according to their strategic importance in a repetitive, intertemporal prisoner's dilemma, cooperation in which may maximize objective income despite the vicissitudes of spontaneous preference. Such a process creates a special realm where discounting is shallow and exponential, a realm that exists as a special case of the matching law in much the same way that Newtonian physics forms a special case of relativistic physics. The consequence for utility theory is that most valuations are made not only or even mainly according to a present hunger, but according to the precedents they will set. If such a valuation process can sometimes make five pounds sterling worth more than ten, it may have the power to account for a number of other anomalies in utility theory as well; but that is another story.

REFERENCES

- Ainslie, G., "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control," *Psychological Bulletin*, No. 4, 1975, 82, 463-96.
- , *Picoeconomics: The Strategic Interaction of Successive Motivational States*, Cambridge: Cambridge University Press, forthcoming.
- and Haendel, V., "The Motives of the Will," in E. Gottheil et al., eds., *Etiology Aspects of Alcohol and Drug Abuse*, Springfield: Charles C. Thomas, 1983.
- Klein, B. and Leffler, K. B., "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*, August 1981, 89, 615-41.
- Lowenstein, G. and Thaler, R. H., "Anomalies: Intertemporal Choice," *Journal of Economic Perspectives*, Fall 1989, 3, 181-93.
- Mazur, J. E., "An Adjusting Procedure for Studying Delayed Reinforcement," in

Economic Analysis and the Psychology of Utility: Applications to Compensation Policy

By DANIEL KAHNEMAN AND RICHARD THALER*

Standard economic analyses often rely on the modern conception of the utility function that resembles a black box. People reveal their utility function through their choices. However, this modern view of utility is not the only possible conception. An earlier view, that associated utility with the pleasures of consumption, prevailed in economics from the days of Daniel Bernoulli and Jeremy Bentham and through the nineteenth century (George Loewenstein, 1991; George Stigler, 1950). We shall refer to this older notion as *experienced utility*, in contrast to the revealed preference notion we call *decision utility*. Since, experienced utility is presumably what people try to maximize and what welfare policies are about, the use of decision utility in economic analyses can only be justified if experienced and decision utility coincide. Recent psychological research, however, suggests several reasons why the concepts might in fact diverge (Kahneman and Carol Varey, 1991; Kahneman and Jackie Snell, 1990).

In this paper we try to illustrate the potential uses of an experienced utility concept by considering some aspects of compensation policy that would normally be ignored in a standard economic analysis. Such factors include the timing of payments over the course of the year, especially the use of lump sum bonuses, and the earnings profile over the worker's life. To a first approximation, the standard economic model assumes that the utility of a compensation package to a worker is invariant with respect to changes in details if the after-tax

present value of the payments is held constant. In this paper we argue for a richer and more psychologically realistic characterization of the quality of experiences that includes the following additional factors: adaptation, contrast, interpersonal comparisons, loss aversion, and fairness. If these factors are included, then it could be possible to create more worker satisfaction without increasing the cost of the pay packages—a free lunch.

I. Adaptation

What is the utility of an income stream? Does the temporal distribution of that stream affect the utility? Can workers be made happier simply by rearranging the timing of their paychecks? These questions are easily dismissed as trivial or meaningless. However, the convergence of evidence from survey data concerning income satisfaction and from psychological studies of adaptation suggests that the questions could be meaningful and the answers interesting. The central problem to be overcome is that of adaptation, the tendency to view that which is perceived as normal to be neutral, neither good nor bad. If people quickly adapt to whatever they are being paid, then no matter how high their salary rises, they never stay happier for long. We first review the evidence that this adaptation process is ubiquitous, and then consider what steps can be taken to mitigate its effects.

The essential results of many studies of income satisfaction and of the effect of income on self-reported happiness are summarized in three propositions (see Michael Argyle, 1987): 1) A cross-sectional study within any society at a particular time is likely to reveal a substantial correlation between income and satisfaction. 2) Societies that differ greatly in income nevertheless have very

*Professor of Psychology, University of California, Berkeley, CA 94720, and H. J. Louis Professor of Economics, Johnson School of Management, Cornell University, Ithaca, NY 14853, respectively. We thank George Loewenstein for helpful comments, and the Russell Sage Foundation for financial support.

similar mean levels of satisfaction. Furthermore, large changes of income level over time within any one society have no effect on income satisfaction. 3) Comparisons to others and especially to one's past determine the standard of satisfaction with income.

These results should probably not be dismissed as a mere artifact of verbal response, because they fit a broader class of observations—about adaptation and habituation to steady states (Harry Helson, 1964). Consider the familiar example of adaptation to temperature. As we all know, there is a substantial "comfort zone" within which we can adapt to temperature. The adaptation temperature is experienced as neutral, or hardly experienced at all. Temperatures that are higher or lower are experienced as hot or cold. To convince yourself of the power of the effect, prepare three pans of water: hot, cold, and tepid. Place one hand in the hot pan, one in the cold one. Wait a short while, then put both hands into the middle pan. One will feel hot, the other cold. This is an example of the phenomenon of adaptation. One of the functions of this mechanism is to maximize the throughput of information by treating a maintained state as a default or null state, and deviations of it as information to be signalled. In the context of hedonic experience (utility, happiness etc.), the phenomenon of adaptation has suggested to some that people are doomed to march forever on a hedonic treadmill (Philip Brickman and Donald Campbell, 1977). This pessimistic conclusion is in accord with the studies mentioned above, concerning the effects of income on welfare. It is also supported by Brickman et al.'s (1978) study of paraplegics and lottery winners. Surprisingly, once these two groups had become adapted to their new situations, they rated their own happiness quite similarly to other people.

It may be possible to defeat or at least retard adaptation, and in so doing, obtain more utils per unit of pay. To do so, however, we must exploit some relevant characteristics of the adaptation mechanisms (the plural is deliberate, because several distinct mechanisms serve the adaptation function).

Here we single out three of them.

1) To enhance the incidence or intensity of a particular sensation, it should be made intermittent over time (and scattered over space, if that dimension is relevant). Putting on red glasses is not a good strategy to obtain an intense experience of seeing red. Indeed, the most intense reds will be perceived in the spatial and temporal vicinity of contrasting greens. Contrast enhances experience. The implications of this point for various hedonic experiences should be obvious.

2) A single experience is less likely to alter the reference level if it is viewed as unusual or distinctive. This result is illustrated by a series of studies in which subjects judge the weight of each of a series of objects lifted, in sequence. In general, the geometric mean of the weights lifted becomes the adaptation level, and is perceived as neither heavy nor light. In some studies, subjects first lift (but do not assess the weight of) some "anchor stimulus" (such as a tray of objects) before every trial. A reliable finding is that the anchor affects the adaptation level, although not to the same extent as it would if it were made relevant to the judgment task. Thus, a heavy anchor will pull up the adaptation level. A more interesting finding is that the anchor has very little effect if it is qualitatively different from the stimuli judged in the series. If the anchor is made very distinctive and ostensibly irrelevant, it has no effect at all. These results suggest that a distinctive pleasurable event will not have a great impact on adaptation level. A bonus need not make the next paycheck appear painfully small.

3) Gradual changes and "spikes" have rather different effects on adaptation levels. Sudden changes are noticed and evaluated as a distinctive departure from adaptation, whereas a very slow gradual change will drag the adaptation level along with it, and may not even be detected.

The general conclusion is that it is certainly possible to use the same amount of money to produce different amounts of utility! The suggestion is that, for an income stream that is sufficient to prevent serious deprivation (compare comfort in the tem-

perature example), there exists a positively skewed distribution of the income over time that will yield greater utility than an even distribution. In particular, taking a portion of the compensation and paying it in a lump sum would appear to make things better.

To see whether this idea appeals to prospective employees, we conducted a small pilot study of attitudes about payment schedules. A group of Cornell University undergraduate students taking a class in psychology were asked whether they would prefer a salary of \$26,000 paid in weekly increments of \$500 or a salary of \$25,000 paid in equal weekly installments and a \$1000 bonus paid midyear. Nearly three-quarters of the subjects (73 percent) preferred the latter plan. (Larger bonuses were less popular.)

II. Side Effects of Bonuses: Saving and Splurging

Our discussion so far has treated the experienced utility of income as a criterion to be maximized. This is somewhat misleading, of course, because the utility of receiving a paycheck and the general sense of satisfaction with one's income are not what most people work for. The evaluation of a payment schedule implicitly invokes assumptions about the scheduling of consumption. Our argument therefore depends on the conjecture that people do not perfectly smooth their consumption when their compensation includes one or two fairly large increments to an otherwise even rate of pay. Instead, they tailor their routine expenses to their "normal" weekly or monthly income, with a different pattern of spending for the lump payments. Our second conjecture is that the altered pattern of expenditures is likely to be advantageous both to the workers and to the economy.

In 1976, Tibor Scitovsky used the principles of adaptation theory to develop an argument about the optimal use of income for consumption. He drew an interesting distinction between comforts, that become noticeable only when they are withdrawn, and pleasures—that are noticeable by definition. An essential condition for pleasure, in Scitovsky's analysis, is contrast with re-

spect to a norm. Feasts, vacations, and gifts are typical occasions for pleasure. Note that the occasion need not be unexpected—deviation from routine suffices. Scitovsky went on to present a rather jaundiced view of the American culture of consumption that appears to favor comforts over pleasures to a greater extent than many others. A significant aspect of the contrast between American and European patterns of consumption involves vacations, that are longer, more elaborate and relatively more costly in Europe. Scitovsky's argument (quite similar to ours above) is that there is a significant sense in which money expended on pleasures is better spent than money spent on comforts. We conjecture that substantial lump sum payments in December and in July would increase spending on pleasures.

Certain bonuses are also likely to increase saving. (There is no inconsistency here. Both pleasurable consumption and saving can increase if routine expenditures decrease. We are predicting that workers would save and splurge more at the expense of a slightly reduced normal standard of living.) A theoretical analysis based on the self-control aspects of saving is provided in Hersh Shefrin and Thaler (1988). They point out that individuals take actions to create their own lumpy payments. For example, a large majority of tax payers receive refunds, thereby giving the IRS an interest-free loan. Empirical support for the effect of bonuses on saving is presented in Tsuneo Ishikawa and Kazuo Ueda (1984). They estimated that in Japan, where most workers receive bonuses twice a year, the marginal propensity to consume bonus income is .437 compared to .685 for regular income. They also separated bonuses into anticipated and unanticipated components, and observed the same marginal propensity to consume from each.

III. Advantages to the Firm

We now turn to possible consequences of our sample proposal for the employing firm. We again concentrate our attention on arguments that would not generally receive much attention from the economics commu-

nity. The topics we consider are stimulating cooperation and increasing wage flexibility.

Stimulating Cooperation. The standard reaction by economists to the alleged incentive effects of profit-sharing schemes is to raise the so-called $1/N$ problem. In a large firm, a worker gets only a trivial portion of any increase in firm profits that he or she may produce, so the incentive to free ride will be strong. While it is possible to construct repeated play economic models in which cooperation is optimal, we agree with David Card's assessment: "I remain sceptical that simple economic models of individual self-interest can usually explain the effects of profit sharing" (1990, p. 141). Of course, the fact that economic theory predicts that profit sharing will have trivial incentive effects does not make it true. This prediction is much like other predictions of free riding. Yet we know that millions of people vote, clean up campgrounds, and donate to public television and other charities. Cooperation in public goods and prisoner's dilemma situations has also been extensively studied in the laboratory by social psychologists and experimental economists in recent years. Two main conclusions emerge from this research (summarized in Robyn Dawes and Thaler, 1988). First, cooperation is common even when the dominant strategy is to defect or free ride. Second, creating a group identity (such as by allowing the members of a group to talk to one another before making anonymous choices) fosters cooperation. It strikes us as plausible that profit-sharing plans can also contribute to the sense of solidarity that would improve worker performance. The empirical evidence gives some support to this view (see Edward Lawler, 1981, and the studies in Alan Blinder, 1990, and Ronald Ehrenberg, 1990).

Fairness and Wage Flexibility. One of the advantages of a Japanese-style bonus plan that Martin Weitzman (1984) has stressed is to create more flexibility in compensation policies and thereby reduce unemployment rates in recessions. Indeed, Richard Freeman and Weitzman (1987) do find that in Japan bonuses are more variable than base pay, and are also more highly

correlated with profits. From a theoretical perspective, it is hard to understand why this should be true. Put another way, why is it necessary to call some of the annual compensation a bonus to achieve greater flexibility? Why should this labeling matter?

Some of our research on perceptions of fairness, done in collaboration with Jack Knetsch (Kahneman et al., 1986), provides a possible answer. Respondents to a telephone survey were asked a series of questions about the perceived fairness of various pricing and wage-setting actions by firms. Three of our findings are germane to the issue of bonuses. 1) Pay cuts by profitable firms even in labor markets with high unemployment are judged to be very unfair. 2) Pay cuts that are occasioned by a deterioration in the competitive position of the employing firm are considered acceptable. 3) Cuts in remuneration are much more likely to be considered fair if they are described as the elimination of bonuses rather than as reductions in the regular wage rate.

The practical suggestion that emerges from this work is that the principle of wage flexibility will be most easily accepted by labor in a context of a gain- (and risk) sharing plan that uses periodic bonuses. If it became policy to encourage a move toward what Weitzman calls a share economy, this bit of psychological wisdom could become useful.

IV. Other Applications: Age Earnings Profiles

Many of the same psychological principles that we have used to argue for bonuses can also be used to explain the anomalous pattern of wages over the life cycle. Several authors have observed that while productivity typically peaks well before retirement, compensation often rises monotonically with age until retirement. Academic salaries are surely a case in point. Edward Lazear (1981) has argued that firms adopt this pattern to reduce shirking by older workers, but Robert Frank and Robert Hutchens (1990) find that the same pattern is observed for airline pilots, where agency problems would seem to be minimal. Frank-Hutchens and Loewenstein and Nachum Sicherman (1991)

both argue that the rising wage profile is better explained by worker preferences.

Adaptation, loss aversion, and self-control are important factors that help explain why workers prefer rising wage profiles. Because workers adapt to current levels of wages, and are particularly sensitive to reductions in their standard of living, a hump-shaped pattern mimicking productivity has obvious drawbacks. The modest positive slope actually observed in real terms, plus the nominal increases necessary to keep up with inflation, can also go a small way toward mitigating the hedonic treadmill. The costs of self-control must be invoked to explain why workers can't transform a hump-shaped wage profile into an increasing consumption profile.

V. Conclusions

Our discussion of the benefits of a bonus system to the workers has been deliberately provocative, raising points that are likely to appear bizarre to many economists, who are accustomed to think within the standard rational model. Our point, of course, is that this may not be the only way to think—and perhaps not even the best. We have made or implied the following assertions: (i) that spending habits will be affected by the temporal distribution of income flows, even when the timing and size of the variations in the flow are largely predictable; (ii) that people will have a preference for a moderate amount of lumpiness; (iii) that people may not be able to pick out the pattern of consumption that would maximize their "real" utility; they may both need and want help in this task.

A possible objection to the arguments we have raised is a familiar one: if a system of bonuses is so good, how come more firms do not use it? In Weitzman's book about the share economy, he responds to this question in part with a joke. How many economists does it take to screw in a light bulb? None. If it were a good idea to screw in the light bulb, someone would already have done it. While many other serious answers can be offered, including cultural and institutional factors, a deeper point may

have to do with the nature of maximizing. It is readily conceivable that a firm may have achieved a local maximum, such that any small adjustment that it makes will be detrimental. This does not imply, however, that there is no larger move that would get to a higher elevation. Consideration of the range of possible organizations of labor and pay suggests that the profit-maximizing firms of the standard model could well be trapped (even if completely successful at the game they are playing) at the peaks of rather lowly hills. This possibility is underscored by the fact that firms in other countries have successfully adopted the system under consideration.

REFERENCES

- Argyle, Michael, *The Psychology of Happiness*, London: Methuen, 1987.
- Blinder, Alan, *Paying for Productivity: A Look at the Evidence*, Washington: Brookings Institution, 1990.
- Brickman, Philip and Campbell, Donald, "Hedonic Relativism and Planning the Good Society," in M. H. Appley, ed., *Social Comparison Processes: Theoretical and Empirical Perspectives*, New York: Wiley/Halsted, 1977.
- _____, Coates, D. and Janoff-Bulman, R., "Lottery Winners and Accident Victims: Is Happiness Relative?," *Journal of Personality and Social Psychology*, August 1978, 36, 917–27.
- Card, David, "Comment," in Alan Blinder, ed., *Paying for Productivity: A Look at the Evidence*, Washington: Brookings Institution, 1990.
- Dawes, Robyn and Thaler, Richard H., "Anomalies: Cooperation," *Journal of Economic Perspectives*, Summer 1988, 2, 187–97.
- Ehrenberg, Ronald, *Do Compensation Policies Matter?*, Ithaca: Industrial and Labor Relations Press, 1990.
- Frank, Robert and Hutchens, Robert, "Feeling Good vs. Feeling Better: A Life-Cycle Theory of Wages," working paper, Cornell University, 1990.
- Freeman, Richard and Weitzman, Martin, "Bonuses and Employment in Japan,"

- Journal of the Japanese and International Economies*, No. 2, 1987, 1, 168-94.
- Helson, Harry, *Adaptation Level Theory*, New York: Harper and Row, 1964.
- Ishikawa, Tsuneo and Ueda, Kazuo, "The Bonus Payment System and Japanese Personal Savings," in M. Aoki ed., *The Economic Analysis of the Japanese Firm*, Amsterdam: North-Holland, 1984.
- Kahneman, Daniel, Knetsch, Jack L. and Thaler, Richard H., "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review*, September 1986, 76, 728-41.
- _____ and Snell, Jackie, "Predicting Utility," in R. Hogarth, ed., *Insights in Decision Making*, Chicago: University of Chicago Press, 1990.
- _____ and Varey, Carol, "Notes on the Psychology of Utility," in J. Romer and J. Elster, eds., *Interpersonal Comparisons of Well-Being*, Cambridge: Cambridge University Press, 1991.
- Lawler, Edward E. III, *Pay and Organization Development*, New York: Addison-Wesley, 1981.
- Lazear, Edward, "Agency, Earnings Profiles, Productivity, and Hours Restrictions," *American Economic Review*, September 1981, 71, 606-20.
- Loewenstein, George, "Intertemporal Choice in Economics," in his and J. Elster, eds., *Intertemporal Choice*, New York: Russell Sage, 1991.
- _____ and Sicherman, Nachum, "Do Workers Prefer Increasing Wage Profiles?" *Journal of Labor Economics*, forthcoming 1991.
- Scitovsky, Tibor, *The Joyless Economy*, Oxford: Oxford University Press, 1976.
- Shefrin, Hersh M. and Thaler, Richard H., "The Behavioral Life-Cycle Hypothesis," *Economic Inquiry*, October 1988, 26, 609-43.
- Stigler, George J., *Five Lectures on Economic Problems*, New York: Macmillan, 1950.
- Weitzman, Martin L., *The Share Economy: Conquering Stagflation*, Cambridge: Harvard University Press, 1984.

Negative Time Preference

By GEORGE LOEWENSTEIN AND DRAZEN PRELEC*

The man who lives within his income is naturally contented with his situation, which, by continual, though small accumulations, is growing better and better every day. He is enabled gradually to relax, both in the rigour of his parsimony and in the severity of his application; and he feels with double satisfaction this gradual increase of ease and enjoyment, from having felt before the hardship which attended the want of them.

Adam Smith
The Theory of Moral Sentiments

Planning for the future invariably requires one to choose among alternative sequences of outcomes. Even simple short-term scheduling decisions about work, play, chores, vacations, etc., involve choosing between sequences, because events that take up time cannot be rescheduled without changing the timing of other activities.

Most economic analyses of preferences between temporally spaced sequences rely on the discounted utility model, along with the assumption of positive time preference and diminishing marginal utility.¹ Barring any preferential interactions across different time periods, the predictions of this model for determining the optimal sequencing of a given set of events are simple: Place the best event at the start, then proceed in descending order until the worst event is

reached at the end. Thus a declining series of consumption levels ought to be preferable to an increasing series, holding total consumption constant.

In this paper, we present a short selection of findings (reported more fully in our 1990 paper) that sharply contradict the normative sequencing rule just described. To most persons, a deteriorating series of utility levels is a rather close approximation to the *least* attractive of all possible patterns, regardless of the nature of events that are being ordered. As a secondary violation of the discounted utility model, the preferences of many people are not additive. Such additivity violations often reflect a concern for spreading utility levels evenly over time that is not attributable to diminishing marginal utility within periods.

I. Sequences vs. Simple Outcomes

Several recent studies have documented an apparently negative rate of time preference for choices among outcome sequences. Loewenstein and N. Sicherman (1991) found that a majority of museum visitors preferred increasing wage profiles over those that are flat or decline over time (holding total value constant). Pointing out that the flat and declining wage profile could produce a *dominating* consumption stream through a suitable savings program, did not have much impact on preference. C. Varey and D. Kahneman (1990) found that subjects strongly preferred brief sequences of decreasing discomfort, even at the cost of experiencing overall greater discomfort, while W. T. Ross and I. Simonson (1990) showed that people prefer sequences that end on a good note.

Preference for improvement is an overdetermined phenomenon, driven in part by anticipatory savoring and dread (Loewenstein, 1987), and in part by loss aver-

*Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15215, and Harvard Graduate School of Business Administration, Boston, MA 02136, respectively.

¹The applications of this model have been remarkably diverse, ranging from savings behavior, educational investment, labor supply, bargaining, to criminal behavior and addiction. In recent years, however, its descriptive validity has come under challenge (see our forthcoming article, and Loewenstein and Richard Thaler, 1990).

sion (A. Tversky and Kahneman, 1990) and habit forming changes in the adaptation level (J. Duesenberry, 1949). Savoring and dread contribute to the preference for improvement because, for gains, improving sequences allow decision makers to savor the best until the end of the sequence. With losses, getting the worst outcomes over with quickly eliminates dread. Adaptation and loss aversion induce preference for improvement because, over time, people tend to assimilate to ongoing stimuli and to evaluate new stimuli relative to their assimilation level. Thus, changes in, rather than levels of, consumption are the carriers of value. Improving sequences afford a continual series of positive departures (gains) from one's adaptation level; declining sequences provide a series of relative losses.

Savoring and dread apply to single outcomes as well as to sequences, but assimilation and loss aversion apply only to sequences. The fact that two motives operate for sequences but only one for simple outcomes suggests that the tendency to defer desirable outcomes will be stronger when those outcomes are embedded in sequences. Such a pattern is illustrated by a survey conducted with undergraduates at Harvard University, who were asked the following three questions:

1. Which would you prefer if both were free?
 - A. Dinner at a fancy French restaurant (86%)
 - B. Dinner at a local Greek restaurant (14%)

For those who prefer French:

2. Which would you prefer?
 - C. Dinner at the French restaurant on Friday in 1 month (80%)
 - D. Dinner at the French restaurant on Friday in 2 months (20%)
3. Which would you prefer?
 - E. Dinner at the French restaurant on Friday in 1 month and dinner at the Greek restaurant on Friday in 2 months (43%)
 - F. Dinner at the Greek restaurant on Friday in 1 month and dinner at the French restaurant on Friday in 2 months (57%)

We anticipated that more people would delay the fancy French dinner when it was combined in a sequence with the Greek dinner than when it was expressed as a single outcome prospect. This is indeed what happened. Of the 86 percent of subjects who preferred the fancy French dinner, 80 percent preferred a more immediate dinner (option C) over a more delayed dinner (option D). However, when the French dinner was composed into a sequence with the Greek dinner, a slight majority (57 percent) preferred to have the better dinner come later. Even with single-outcome events there is some motivation to defer the French dinner—witness the 20 percent of subjects who opted for the longer delay. However, this tendency is stronger for sequences than for individual items.

A similar pattern is observed when "Dinner at home" is substituted for the Greek dinner. Since most people eat dinner at home on most nights anyway, embedding the French dinner in an explicit binary sequence does not introduce any real modification of the problem, but the subject is reminded that the choice is "really" between complete sequences. Like other framing effects, such reminders cause preferences to shift, in this case in favor of the improving sequence.

II. What Is a Sequence?

If impatience and the desire for improvement are simultaneously present within a single individual, what determines the prepotent motive on a given occasion? A reasonable conjecture is that the desire for improvement depends on the "integrity" of the sequence—the extent to which the events that comprise it are of a similar type, are regularly spaced, and are not stretched too far apart.

The following example illustrates how the integrity of sequence can be reduced in a predictable way. Visitors to a science museum were asked to choose between alternative scheduling of two visits to a city where the respondent once lived, one to be spent with "an irritating, abrasive, aunt who is a horrendous cook," the other with "former

work associates whom you like a lot." Each subject made three choices, as described below. Response frequencies are reported in parentheses.

	This weekend	Next weekend	
A.	friends	aunt	(10%)
B.	aunt	friends	(90%)
	This weekend	26 weeks from now	
C.	friends	aunt	(48%)
D.	aunt	friends	(52%)
	26 weeks from now	27 weeks from now	
E.	friends	aunt	(17%)
F.	aunt	friends	(83%)

Observe that it is not possible to interpret the three modal choice patterns as the result of a *single* time preference function, denominated in absolute time. Responses to the first and third questions suggest a negative rate of time preference over a 1-week interval, irrespective of whether the week is close or far. By interpolation, one may conclude that the aunt would be scheduled in the earlier of *any* consecutive 2 weekends. Yet, for many subjects, the negative time preference over adjacent weekends does not compound into a negative time preference for the complete 6-month interval, as shown by responses to the second question.

Applied uniformly to all choices, negative time preference would require harsh reductions in present consumption in favor of the future. The fact that one does not observe such sacrifices, even given the additional inducements of a positive interest rate, is normally taken as evidence for positive time preference (M. Olson and M. J. Bailey, 1981).

We propose a different interpretation: Negative time preference is applied selectively, to those events that are seen as part of a meaningful sequence, having a well-defined starting and ending point. As the previous example shows, a pair of adjacent weekends defines a minimal but coherent

sequence, while a pair of weekends separated by 6 months does not.

The saliency of particular intervals is not an objective matter, but depends rather on perceptual framing. By deliberately manipulating the subjective frame, it is possible to induce normatively peculiar patterns of choices, as the following example shows. Subjects were asked to indicate their preferred times to eat two free dinners at the restaurant of their choice. One group was given no time constraint concerning when they could eat the dinners. A second, *constrained* group was told that the dinners must be consumed within the next 2 years. We anticipated that constrained subjects would prefer to delay the dinners more than unconstrained subjects, because the introduction of an explicit planning interval reminds the subjects that, by selecting a particular pair of dates, they are also choosing *not to consume* the meals on all of the other dates. As predicted, the mean preferred delay for the first dinner was 3.3 weeks for the unconstrained group and 7.7 for the constrained. Mean preferred delay times for the second dinner were 13.1 and 31.1 weeks.

These results are inconsistent with the axiom of revealed preference. The imposition of a time constraint on an initially unconstrained population should only affect the responses of that fraction of the population whose preferred delays are longer than permitted by the constraint. The population averages should, therefore, be longer in the unconstrained condition.

III. Nonadditive Preferences

Aside from a preference for improvement, choices between sequences also reveal a sensitivity to certain global or "gestalt" properties, having to do with how evenly the good and bad outcomes are arranged over the total time interval. Consider the following problem presented to 37 Yale University undergraduates (from Loewenstein, 1987). Subjects were first given a choice between A and B, then between C and D. Percentages choosing each of the options are presented in the right-hand column.

Alternative	Weekend 1	Weekend 2	Weekend 3	Choices
A	Fancy French	Eat at home	Eat at home	16
B	Eat at home	Fancy French	Eat at home	84
C	Fancy French	Eat at home	Fancy Lobster	54
D	Eat at home	Fancy French	Fancy Lobster	46

In the first problem, the majority of subjects preferred to postpone the fancy dinner to weekend 2, in keeping with the improvement principle. However, in the second problem, the insertion of the common lobster dinner in weekend 3 caused preference to shift slightly in favor of having the French dinner right away. This is a violation of additive separability, which implies that an individual preferring sequence A over B should continue to prefer A over B if any elements shared by the two sequences are altered in the same way.

Intertemporal additivity has never been viewed as normatively compelling, since there are many situations in which it is reasonable for consumption at one point in time to influence the marginal utility of consumption at another. Models of the "habit formation" type (Duesenberry; R. A. Pollack, 1970; G. M. Constantinides, 1990) assume that instantaneous utility depends negatively on past consumption. Other models incorporate the rate of consumption change into the utility function (R. Frank, 1989), or a preference/aversion for utility variation between adjacent periods, as in I. Gilboa's elegant formulation (1989).

Can the separability violation described above be reasonably attributed to some sort of loss aversion, following an adaptation to a reference point? An aversion to utility *reductions* from one period to the next would seem to work in favor of alternative D over C, while leaving preferences between A and B unchanged (to a first approximation). Hence it would predict the opposite violation pattern to the one actually observed.

Furthermore, we have observed the same preference pattern when common "Eat at home" weekends are inserted between the original weekends 1 and 2, and 2 and 3, in all four alternatives (thereby creating 5-weekend sequences). Because the neutral

filler weekends should reduce adaptation, and altogether eliminate differential interactions between adjacent periods, they should attenuate the separability violations. The results with the 5-weekend version of the problem were essentially equivalent, however: Only 11 percent of respondents opted for option A over B, while 49 percent preferred C over D.

The problem here is that habit formation/loss aversion models do not well capture the global properties that people find attractive in sequences. The relative advantage of sequences B and C is due to the fact that they both "cover" the 3-week interval better than their respective alternatives. In other words, they interleave the good and indifferent events in a more nearly uniform manner.

We have developed a theoretical model that measures both the degree of improvement and spreading of any sequence in terms of cumulative utility sequences (see our 1990 paper). The model defines improvement over time as the sum of deviations of the cumulative sequence that would be obtained by spreading total utility evenly over time from the cumulative utility stream of the sequence being evaluated. Evenness of spread is represented by the sum of the absolute value of these deviations.

The derivation of our notions of improvement and uniformity is depicted in Table 1, using the last illustrative example on the assumption that the "Eat at home" event has utility zero, and the "Fancy French" and "Fancy Lobster" events have utility one. Note that a simple preference for improvement would lead to a preference for B over A and D over C, while discounting alone would produce the opposite pattern. The desire for spreading outcomes over time, however, designated in the row marked Spread (lower numbers signify more even spreading of outcomes), can explain the preference for B over A and for C over D.

In several longer surveys (see our 1990 paper), we have systematically mapped out preferences over multiple period sequences. The judgments of the average person could be briefly described as follows: There is a strong liking for improving sequences, mod-

TABLE 1—DERIVATION OF IMPROVEMENT AND SPREAD MEASURES

	A	Alternative B	C	D
Sequence	1, 0, 0	0, 1, 0	1, 0, 1	0, 1, 1
Cumulative	1, 1, 1	0, 1, 1	1, 1, 2	0, 1, 2
Flat				
Seq.	.33, .33, .33	.33, .33, .33	.66, .66, .66	.66, .66, .66
Flat				
Cumul.	.33, .66, 1	.33, .66, 1	.66, 1.33, 2	.66, 1.33, 2
Difference	-.66, -.33, 0	.33, -.33, 0	-.33, .33, 0	.66, .33, 0
Improve- ment	-1	0	0	1
Spread	1	.66	.66	1

erated by a penalty for deviation from global uniformness, and a small premium for sequences that start well.

IV. Conclusion

Previous psychological work on time preference has focused almost entirely on the tradeoff that arises when two outcomes of different dates and different values are compared. The tacit premise was that such judgments will reveal an individual's "raw" time preference, from which one can then synthesize preferences over more complex objects—retirement plans, intertemporal income profiles, and such. This view we feel is fundamentally incorrect: As soon as an intertemporal tradeoff is embedded in the context of two alternative *sequences* of outcomes, the psychological perspective, or "frame" shifts, and individuals become more farsighted, usually wishing to postpone the better outcome to the end. The same person who prefers a good dinner sooner rather than later, if given a choice between two explicitly formulated sequences, one consisting of a good dinner *followed* by an indifferent one, the other of the indifferent dinner *followed* by the good one, may well prefer the latter alternative. Sequences of outcomes that decline in value are greatly disliked, indicating a negative rate of time preference.

A byproduct of the sequence frame is that subjects who are given a time interval, within which to schedule some enjoyable activity, may schedule it later on average

than people who are given no time frame at all. Apparently, as soon as the relevant interval is specified, a person becomes concerned with shifting the good events out to the end. This result has implications for life cycle choices; for example, it suggests the possibility that some individuals would choose an earlier retirement in the absence of a mandated retirement point.

The sensitivity of time preference to the sequence "frame" casts new light on the often-repeated charge that certain groups of people (consumers, managers, members of a particular nation or culture) have an excessively steep rate of time preference. Such a claim is a psychologically imprecise definition of the problem, at best. The differences that do prevail should instead perhaps be traced to different styles of mental book-keeping, which will alone produce different degrees of impatience even with a common underlying rate of time preference. Any operation, custom, or habit that causes the stream of purposeful activity to fragment into a series of isolated choices, each involving a simple intertemporal tradeoff, and each unrelated to a larger plan, encourages impatient choices. Whereas the integral sequence frame, by fusing events into a coherent sequence, promotes concern for the future, thereby creating an appearance of negative time preference.

REFERENCES

- Constantinides, G. M., "Habit Formation: A Resolution of the Equity Premium Puzzle," *Journal of Political Economy*, June 1990, 98, 519–43.
- Duesenberry, J., *Income, Saving, and the Theory of Consumer Behavior*, Cambridge: Harvard University Press, 1949.
- Frank, R., "Frames of Reference and the Quality of Life," *American Economic Review Proceedings*, May 1989, 79, 80–85.
- Gilboa, I., "Expectation and Variation in Multi-Period Decisions," *Econometrica*, September 1989, 57, 1153–69.
- Loewenstein, G., "Anticipation and the Valuation of Delayed Consumption," *Economic Journal*, September 1987, 97, 666–84.

- _____ and Prelec, D., "Anomalies in Intertemporal Choice: Evidence and an Interpretation," *Quarterly Journal of Economics*, forthcoming.
- _____ and _____, "Preferences over Outcome Sequences," Harvard Business School Working Paper, 1990.
- _____ and Sicherman, N., "Do Workers Prefer Increasing Wage Profiles?," *Journal of Labor Economics*, January 1991, 9.
- _____ and Thaler, R., "Anomalies: Intertemporal Choice," *Journal of Economic Perspectives*, Fall 1989, 3, 181-93.
- Pollak, R. A., "Habit Formation and Dynamic Demand Functions," *Journal of Political Economy*, July/August 1970, 78, 745-63.
- Olson, M. and Bailey, M. J., "Positive Time Preference," *Journal of Political Economy*, February 1981, 89, 1-25.
- Ross, W. T., Jr. and Simonson, I., "Consumers' Evaluation of Purchase and Consumption Experiences: A Preference for Happy Endings," unpublished manuscript, 1990.
- Tversky, A. and Kahneman, D., "Reference Theory of Choice and Exchange," unpublished working paper, 1990.
- Varey, C. and Kahneman, D., "The Integration of Aversive Experiences Over Time: Normative Considerations and Lay Intuitions," working paper, University of California-Berkeley, 1990.

Designing Economic Agents that Act Like Human Agents: A Behavioral Approach to Bounded Rationality

By W. BRIAN ARTHUR*

Most economists accept that there are limits to the reasoning abilities of human beings—that human rationality is bounded. The question is how to model economic choices made under these limits. Where, between perfect rationality and its complete absence, are we to set the “dial of rationality,” and how do we build this dial setting in to our theoretical models?

One approach to this problem is to lay down axioms or assumptions that suppose limits to economic agents’ computational ability or memory, and investigate their consequences. This is useful, but it begs the question of how humans actually behave. A different approach (the one I suggest here) is to develop theoretical economic agents that act and choose in the way *actual* humans do. We could do this by representing agents as using parametrized decision algorithms, and choose and calibrate these algorithms so that the agents’ behavior matches real human behavior observed in the same decision context. Theoretical models using these “calibrated agents” would then, we could claim, furnish predictions based on *actual* rather than idealized behavior.

It is unlikely there exists some yet-to-be-defined decision algorithm, some “model of man,” that would represent human behav-

ior in all economic problems—an algorithm whose parameters would constitute universal constants of human behavior. Different *contexts* of decision making in the economy call for different actions; and an algorithm calibrated to reproduce human learning in a search problem might differ from one that reproduces strategic-choice behavior. We would likely need a repertoire of calibrated algorithms to cover the various contexts that might arise.

Nevertheless, for a particular context of decision making, calibrating theoretical behavior to match human behavior would allow us to ask questions that are not answerable at present under the assumption of either perfect rationality or idealized learning. We might want to know whether a given neoclassical model with human agents represented by “calibrated agents” will result in some standard asymptotic pattern—a rational-expectations equilibrium, say. We might ask whether agents calibrated to learn as humans do converge to some form of optimality, or interactively to a Nash equilibrium.¹ And we might want to study the speed of adaptation in a particular economic model with human agents represented by calibrated agents.

What would it mean to calibrate an algorithm to “reproduce” human behavior? The object would be algorithmic behavior that reproduces statistically the characteristics of human choice, including the distinctive errors or departures from rationality that hu-

[†]*Discussants:* Ken Binmore, University of Michigan; Drew Fudenberg, MIT; John Geanakoplos, Yale University.

*F.R.I. Stanford University, Stanford, CA 94305-6084, and Santa Fe Institute, 1120 Canyon Road, Santa Fe, NM 87501. I thank Kenneth Arrow, Frank Hahn, John Holland, Richard Herrnstein, and David Lane for stimulating and discussing these ideas, without implicating them in the views expressed here, and Drew Fudenberg and John Geanakoplos for helpful criticism.

¹Drew Fudenberg and David Kreps (1988) and Paul Milgrom and John Roberts (1991) take a different, but parallel approach. They show that if human learning behavior fulfills certain axioms, asymptotic behavior will result in standard outcomes.

mans make, in the given context. Ideally, the algorithm could pass a Turing test of being indistinguishable from corresponding human behavior in the same context, to an observer who was not informed whether the behavior was algorithm generated or human generated (Alan Turing, 1956). This of course would be asking a lot.

This paper reports on and discusses my recent work (1990) that explores this idea of calibrating an algorithm to reproduce human behavior. It develops and calibrates a learning algorithm for a commonly encountered and simple decision context, that of agents choosing repeatedly among discrete actions with initially unknown, random consequences.

I. A Parametrized Learning Automaton

Consider the problem of iterated choice under uncertainty, in which a decision maker chooses one of N possible actions at each time that have random payoffs or profits drawn from a stationary distribution that is unknown in advance. This would be the case, for example, where a firm, government agency, or research department is faced each period with a choice among N alternative pricing schemes, or policy options, or research projects, each with consequences that are poorly understood at the outset and that vary from "trial" to "trial". The agent chooses one alternative at each time, observes its consequence or payoff, and over time updates his choice as a result. What makes this iterated choice problem interesting is the tension between *exploitation* of high-payoff actions that have been undertaken many times and are therefore well understood, and *exploration* of seldom-tried actions that potentially may have higher average payoff.

The classic multi-arm-bandit version of this problem is to design a learning algorithm or automaton that maximizes some criterion—such as expected average payoff. Our problem is different. It is to design a learning algorithm or learning automaton that can be tuned to choose actions in this iterated choice situation the way humans would.

Consider then a learning automaton that represents a single agent who can undertake one action of N possible actions at each time. We may think of "learning" in this iterated-choice context as updating the probabilities of taking each action on the basis of the payoffs or outcomes experienced. Action i brings reward $\Phi(i)$ that is unknown to the agent in advance, positive, and distributed randomly with a stationary distribution. The automaton (our artificial agent) "learns" via the following simple algorithm. It associates a vector of *strengths*, S_t , with the actions 1 through N , at each time t . The current sum of these strengths is C_t (the component sum of S_t), and the initial strength vector S_0 is strictly positive. The vector p_t represents the agent's probabilities of taking actions 1 through N at time t . At each time, t , the agent:

- 1) Calculates the probability vector as the relative strengths associated with each action. That is, it sets $p_t = S_t / C_t$.

- 2) Chooses one action from the set according to the probabilities p_t and triggers that action.

- 3) Observes the payoff received and updates strengths by adding the chosen action's j 's payoff to action j 's strength. That is, where action j is chosen, it sets the strengths to $S_t + \beta_t$, where $\beta_t = \Phi(j)e_j$; (e_j is the j th unit vector).

- 4) Renormalizes the strengths to sum to a value from a prechosen time sequence. In this case, it renormalizes strengths to sum to $C_t = Ct$.

This last step allows us to set the rate and deceleration of the learning via the parameters C and ν that are fixed in advance. The rate of learning, it turns out, is proportional to $1/(Ct)$. Parameters C and ν thus define a two-parameter family of algorithms that can be used to calibrate the automaton.

The algorithm has a simple behavioral interpretation (at least when $\nu = 0$). The strength vector summarizes the current confidence the agent or automaton has learned to associate with actions 1 through N . Confidence associated with an action increases according to the (random) payoff it brings in when taken. The automaton chooses its action with probabilities proportional to its

current confidence in the N actions, and learning takes place as these probabilities of actions are updated. The summed confidence in all actions is constrained to be constant. S_0 , the initial confidence in the actions, represents prior beliefs, possibly carried over from past experience.

It also has a machine-learning interpretation. A Holland-type *classifier* is a condition/action couple ("if object appears in left vision field/turn toward object"), where the action is allowed to be activated only if the condition is fulfilled (John Holland et al., 1987). If several classifiers have the same condition and that condition is fulfilled, they "compete" to be the one activated. Our algorithm can be viewed as a set of N classifiers each competing to be activated, where classifier j is the simple couple "if it is time to act/choose alternative j ." As is standard in classifier systems, strengths are associated with the classifiers; one classifier is triggered on the basis of current strengths; and the chosen classifier's strength is updated by the associated reward.

The algorithm is nonlinear in that actions that are frequently taken are further strengthened or reinforced, as in the classic Hebb's rule (Donald Hebb, 1949). And it is stochastic in that actions are triggered randomly on the basis of current probabilities, and rewards are drawn randomly from a distribution. Nonlinearity allows for the exploitation of "useful" actions—ones that pay well tend to be strengthened early on and therefore to be heavily emphasized. And the stochastic property (triggering actions randomly on the basis of their strength) allows for exploration: if a little used action brings in a "jackpot," it may be strengthened sufficiently to become a frequent action.

What can we say about the long-run properties of the learning implicit in this algorithm? Will it "discover" the maximal expected-payoff action, k say, and learn over time to activate it only in the limit? This is not obvious. There are two contradictory tendencies. On the one hand, if an inferior high-payoff action j is triggered early, it may gain in strength and be triggered ever more often until it dominates. Learning may

then fall into action j 's "gravitational orbit" without escaping. On the other hand, if exploration does not die away too fast, the algorithm will eventually uncover the fact that k is better and home in on it.

In my earlier paper (1990), I show that the algorithm has stochastic dynamics:

$$(1) \quad p_{t+1}(i) = p_t(i) + \alpha_t \left\{ p_t(i) \left[\phi(i) - \sum_j \phi(j) p_t(j) \right] + \xi_t \right\}.$$

The probability of choosing action i grows at a rate proportional to the difference between i 's expected payoff $\phi(i)$ and the average payoff at current probabilities, plus an unbiased perturbation term ξ . The step-size α_t is $1/(Ct^\nu)$. Further analysis settles optimality. If $\nu < 1$, the step-size remains large enough for inferior action j , emphasized early by chance, possibly to build up sufficient strength to shut k out. Optimality is in this case not guaranteed. If, on the other hand, $\nu = 1$, optimality is guaranteed. The step-size falls off at rate $1/t$; this delays movement to a nonoptimal action and retains exploration for a long enough time for k to be repeatedly activated and to dominate.

II. Calibration Against Human Subjects

We now want to calibrate the parameters C and ν against data on human learning. Here we are interested in three things: the degree to which the calibrated algorithm represents human behavior; whether the measured value of ν lies within the range that guarantees asymptotically optimal choices; and the general characteristics of learning (such as speed and ability to discriminate) that the calibrated values imply.

To calibrate the algorithm, I use the results of a series of two-choice bandit experiments conducted by Laval Robillard at Harvard in 1952–53 using students as subjects (reported in Robert Bush and Frederick Mosteller, 1955). I would prefer to calibrate on more recent experiments, but these have gone out of fashion among psycholo-

gists, and no recent, more definitive results appear to be available. I therefore use Robillard's data as an expedient, interpreting the resulting calibration as a good indication of human behavior in situations of choice rather than a definitive statement.

Robillard set up seven experiments, each with its own payoff structure, and allocated groups of ten subjects to each. Each subject could choose action *A* or *B* repeatedly in 100 trials; and the experiments differed in the probabilities with which unit payoffs occurred. For each experiment, Robillard reported the proportion of *A* choices in each sequential block of 10 trials, averaged over the group of ten subjects (for the data, see my 1990 paper).

I proceed by allowing "groups of artificial agents" (computer runs of the algorithm) to reproduce the equivalent of Robillard's data for fixed values of *C* and ν . The artificial agents produce *stochastic* sequences of choices or frequencies of choosing action *A*; hence goodness of fit to Robillard's data (under a suitable criterion) for fixed parameters is a random variable. I calibrate the parameters *C* and ν by minimizing the expected sum of errors squared between the automata-generated frequencies and the corresponding human frequencies for each particular experiment, totaled over the seven experiments. This results in $C = 31.1$ and $\nu = 0.00$. Note immediately that ν lies in a region where asymptotic optimality is far from guaranteed.

Figures 1 and 2 show the artificial agents' learning plotted against the human subjects' in four of the seven experiments, using these calibrated values. (The other experiments are similar in fit.) Judged by eye, the results are encouraging. The automata learn with roughly the same rate and variation as the humans in each of the experiments. Further statistical work (see my 1990 paper) shows that other data sets besides Robillard's produce similar fits, and that six of the seven Robillard learning trajectories could have been produced by the calibrated automata in the sense that each fits well within a distribution of 100 corresponding computed automata trajectories. (The outlier experiment, pictured second in Figure 2, has

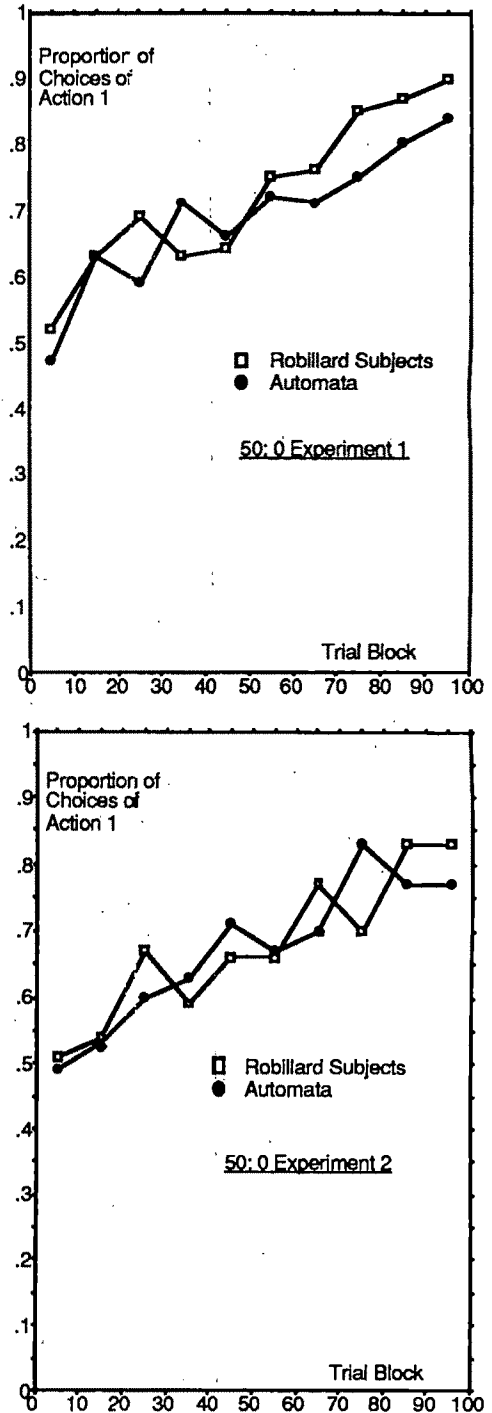


FIGURE 1. CALIBRATED AGENTS' VS. HUMANS' CHOICE FREQUENCIES IN TWO EXPERIMENTS

Note: 50:0 denotes unit payoff with probabilities 50 and 0 percent.

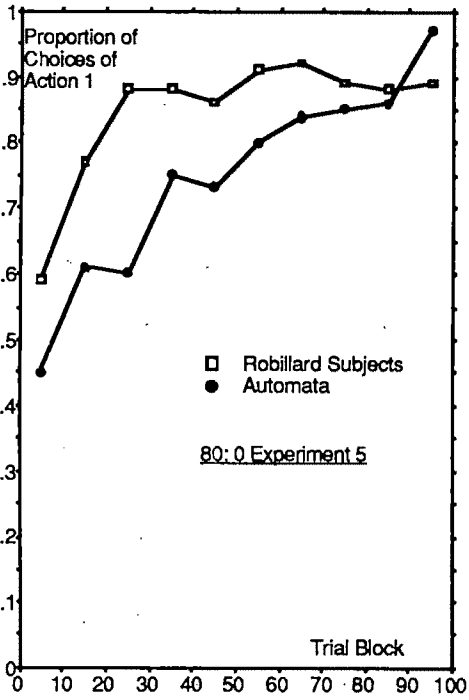
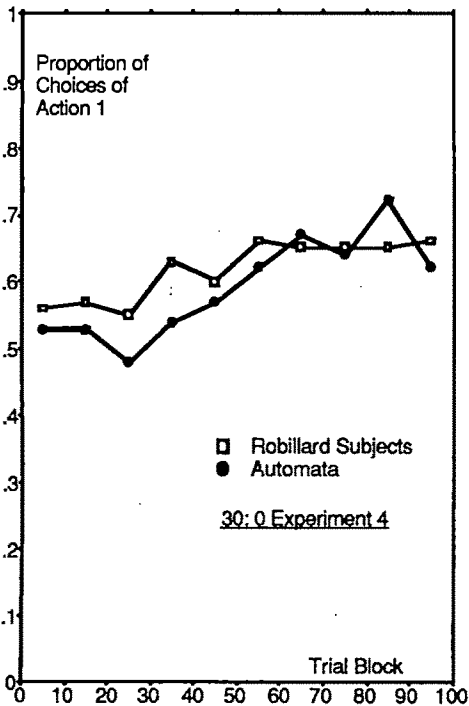


FIGURE 2. CALIBRATED AGENTS VS. HUMANS IN TWO OTHER EXPERIMENTS

close-to-determinate payoffs, something humans notice faster than the algorithm.)

More convincing than statistical tests-of-fit are tests of whether the algorithm can replicate human behavior in quite different choice problems than those for which it was calibrated. A different problem is provided by recent experiments of Richard Herrnstein et al. (1990) where the distribution of payoffs to a choice is no longer fixed, but depends instead on the frequency of actions taken. These are of interest because in this frequency-dependent case, human subjects show a well-documented characteristic behavior called *melioration*; their choices converge not to optimal frequencies that maximize expected payoff, but to quite different frequencies that equalize expected payoffs of each action. In reproducing these experiments, I found that the calibrated agents perform similarly to the Herrnstein subjects; they also meliorate. That our algorithm picks up this characteristic of human behavior is not surprising. Both human and artificial agents carry out local search on the efficacy of choices at the current frequency of choice; thus in this frequency-dependent case both deviate from "rational" behavior in the same way. Findings like this give us confidence that we *can* indeed replicate human behavior for particular decision contexts with calibrated learning agents.

III. What About Optimality and Nash Equilibria?

Let us now put our calibrated agents to work. First, what they can tell us about the prospects for human choices converging to long-run optimality (in our standard, frequency-independent case)? Theoretically, we know from the zero-calibrated value for ν that optimality may not be reached on all occasions. But how often might this happen in practice?

To explore this, I set up a series of computer experiments (see my 1990 paper) designed around an iterated decision problem with six choices, each with uniformly distributed payoff from 0.5 to 1.5 times the choice's expected value. Expected payoffs fell off geometrically from action 1 to 6,

with action 1 the maximal choice. I found that as long as the best action was set to be more than 15 percent better than the next best, the algorithm locked in to it close to 100 percent of the time, in 100 repeated experiments. But when the best action's expected payoff was set at less than 10 percent or so more than the next best's, chance variations in received payoff caused choices to lock in to less-than-optimal outcomes in a significant percentage of experiments. Human choice, if captured by the calibration, appears to "discover" and exploit the optimal action with high probability, *as long as it is not difficult to discriminate*. But beyond a perceptual threshold, where differences in alternatives become less pronounced, nonoptimal outcomes become more likely.

It might be objected that this finding is merely an artifact of the algorithm I have chosen. I do not believe so. What is crucial to the emergence of the optimal action is a slowing down in speed of convergence, so that learning has time to "discover" the action with largest expected value. The data, not the algorithm, show this slowing down does not occur. I would thus expect the finding that long-run optimality depends on the difficulty of the problem to be validated under other well-fitting algorithm specifications. We could of course invoke an imagined discount rate to render lock-in to an inferior outcome "optimal over time." But this "discount rate" would then appear to be independent of the time between trials, and I find this argument unpersuasive.

A similar finding carries over to the question of whether human agents are likely to converge to a Nash equilibrium in an iterated game. Think now of our calibrated agents representing human agents learning within a normal-form, stage game (see my 1989 paper). The agents can observe their own actions and random payoffs, but are not particularly well informed of other players' actions and payoff functions. An example might be oligopolistic firms choosing among pricing policies in a decentralized market on the basis of observed end-of-quarter profit. Each agent then faces a multichoice bandit problem as before, and our

learning context carries over to this wider problem. Of course, in this case, each agent's payoff distribution changes slowly as other agents change their choice probabilities.

We can represent strategic learning here, for each agent separately, by the calibrated stochastic process in equation (1) and apply asymptotic probabilistic analysis to the resulting stochastic, dynamic model. Our results then represent human behavior in this context to the degree that the calibration captures actual human learning.

For some game-payoff structures, it turns out, strategies may not converge at all. The fact that each agent changes his choice probabilities (strategy profile) as other agents change *theirs* may cause strategy profiles to cycle. In games where learning *does* converge, the analysis shows a Nash outcome is likely but not assured. Nash requires that each agent converge to best reply; but with $\nu = 0$, there may not be sufficient exploration of strategies, and Nash is not guaranteed. In practice, as before, the likelihood of convergence to Nash depends on the difficulty of discrimination among the action payoffs.

How might we use calibrated agents to represent actual human adaptive behavior in other standard neoclassical models? My paper in progress with Holland, Palmer, and Tayler explores convergence to rational expectations equilibrium using calibrated agents in an adaptive version of the Lucas (1978) stock market. We find that the calibrated agents learn to buy and sell stock appropriately, and that the stock price indeed converges to small fluctuations around the rational expectations value. However, we also find that speculative bubbles and crashes occur—a hint that under realistic learning technical analysis may emerge.

IV. Conclusion

I conclude from this exploratory exercise that we can indeed design artificial learning agents and calibrate their "rationality" to replicate human behavior. Not only does the learning behavior of our calibrated agents vary in the way human behavior

varies as payoffs change from experiment to experiment in this repeated multichoice context, but it also reproduces two stylized facts of human learning well-known to psychologists: that with frequency-dependent payoffs, humans meliorate rather than optimize; and there is a threshold in discrimination among payoffs below which humans may lock in to suboptimal choices. Most usefully perhaps, the calibrated algorithm has a convenient dynamic representation that can be inserted into theoretical models.

To the degree that the algorithm replicates human behavior, it indicates that human learning most often adapts its way to an optimal steady state or, interactively, to a Nash outcome. But it also shows that humans systematically underexplore less-known alternatives, so that learning may sometimes lock in to an inferior choice when payoffs to choices are closely clustered, random, and difficult to discriminate among. Thus the question of whether human learning adapts its way to standard economic equilibria depends on the perceptual difficulty of the problem itself.

For choices among actions with initially unknown, random payoffs, it appears that behavior does not settle down much before 40 to 100 or more trials. This implies that there is a *characteristic learning time* for human decisions in the economy that depends both on the payoff structure of the problem and on the frequency of observed feedback on actions taken. There is also a time horizon over which the economic environment of a decision problem stays relatively constant. For some parts of the economy, the learning time may be shorter than the problem time horizon. These would be at equilibrium—albeit a slowly changing one. For other parts, learning may take

place more slowly than the rate at which the problem shifts. These parts would be always transient, always tracking changes in their decision environment, and never at equilibrium.

REFERENCES

- Arthur, W. Brian, "A Learning Algorithm that Mimics Human Learning," Santa Fe Institute Working Paper 90-026, 1990.
- _____, "Nash-Discovering Automata for Finite-Action Games," mimeo., Santa Fe Institute, 1989.
- _____, et al., "A Stock Market with Artificially Intelligent Agents," paper in progress, Santa Fe Institute, 1991.
- Bush, Robert and Mosteller, Frederick, *Stochastic Models for Learning*, New York: Wiley & Sons, 1955.
- Fudenberg, Drew and Kreps, David M., "Learning, Experimentation, and Equilibrium in Games," mimeo., MIT, 1988.
- Hebb, Donald O., *The Organization of Behavior*, New York: Wiley & Sons, 1949.
- Herrnstein, Richard et al., "Maximization and Melioration," mimeo., Department of Psychology, Harvard University, 1990.
- Holland, John H. et al., *Induction: Processes of Inference, Learning, and Discovery*, Cambridge: MIT Press, 1987.
- Lucas, Robert E., "Asset Prices in an Exchange Economy," *Econometrica*, November 1978, 46, 1429–45.
- Milgrom, Paul and Roberts, John, "Adaptive and Sophisticated Learning in Repeated Normal Form Games," *Games and Economic Behavior*, February 1991, 3.
- Turing, Alan M., "Can a Machine Think?," in John R. Newman, ed., *The World of Mathematics*, Vol. 4, New York: Simon and Schuster, 1956, 2009–2123.

Experiments on Stable Suboptimality in Individual Behavior

By R. J. HERRNSTEIN*

In several recent experiments,¹ subjects made choices in a way that supports the idea that choice is governed, either sometimes or always, by a principle that does not necessarily maximize utility, as the subjects themselves would reckon their utility. In other words, they behaved irrationally, and their irrationalities seemed to be systematic and motivated, not just a matter of behaving carelessly or erroneously.

I. Representative Data

The subjects in the experiments worked for money by repeatedly choosing one of two alternatives in procedures in which the long-run pattern of choice interacted with the rate of payment for each alternative. The subjects earned less money than they could have, and they did so when there was no plausible compensating nonmonetary gain. Their suboptimality was often consistent with a principle of choice well established by hundreds of experiments mainly on animal behavior (Michael Davison and Diane McCarthy, 1988, Ben Williams, 1988; for related experiments on human subjects, see Christopher Bradshaw and Elemer Szabadi, 1988).

Subjects (mostly students) were recruited at the University of Chicago and Harvard University with posters announcing the possibility of earning a few dollars in a brief

session studying decision making. A subject sat in front of a computer monitor in a small, quiet chamber. The instructions explained that striking the left or right arrow on the keyboard would release a coin from one or the other of the two boxes portrayed on the screen. Each subject made about 300–400 choices of the left or right key for money, after about 100–200 practice choices under the same conditions. Striking the left or right arrow key always earned a coin worth one cent each with a probability of 1.0. The subject was told that payment would be equal to the total value of the coins earned, and also that it was a good idea to try to earn coins rapidly. From time to time, a message would appear on the monitor telling the subject how much time was left until the end of the session.

Figure 1 shows the contingencies of reward in experiment 1.² It shows the time taken for a coin to fall from the right and left boxes (R and L, respectively), as a function of preceding 40 choices that were of the right key. The dotted line is the average time weighted by the abscissa value. The open circles plot individual subjects and the right angle cross shows the average subject. On the vertical axis is the time (in seconds) for a coin to drop into reservoirs from either the left or right box. While one coin is dropping, no new coin could be earned. A session comprised $6\frac{2}{3}$ minutes of practice followed by 15 minutes earning money to keep.³ Coins fell at a speed that depended on the subject's recent choices.

The abscissa value is the proportion of right key choices in the subject's preceding 40 choices, hence updated after every choice. The time it took for a coin to fall depended on where along the x-axis the

*Department of Psychology, Harvard University, Cambridge, MA 02138. All of the experiments to be described grew out of an experimental design first used in the cited work, but much extended in discussions at the Russell Sage Foundation during 1988–89, with Ronald Heiner, George F. Loewenstein, Drazen Prelec, Howard Rachlin, and William Vaughan, Jr. The research was itself supported by the Russell Sage Foundation. The collaborators for particular procedures will be noted in turn.

¹See my paper with Drazen Prelec and William Vaughan (1986).

²This experiment was done by Vaughan, and myself.

³These times refer to elapsed times while coins were dropping, not continuous clock time.

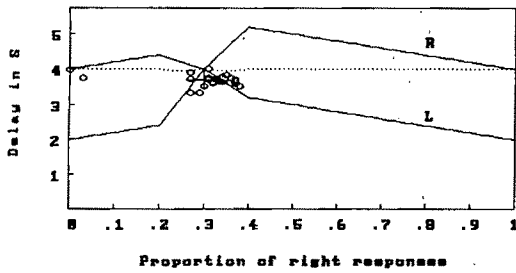


FIGURE 1. EXPERIMENT 1

choice was made. Suppose the preceding 40 choices had been evenly divided between left and right keys and the subject chose the right key. It would take the coin almost 5 seconds to drop. A coin from the left box would have fallen in almost 3 seconds. Indeed, so long as the proportion of right choices in the preceding 40 trials is above .4, coins from the right box take 2 more seconds to fall than coins from the left. Below a proportion of .2, coins from the left box take 2 more seconds to fall than coins from the right. At a proportion of .3 choices of the right key, coins from either box take 4 seconds to fall.

The reward contingencies were chosen so that the weighted average time to drop, traced by the dotted line, was 4 seconds, no matter what the subject did. Consider the two extremes and the midpoint. At the midpoint, half the coins were taking about 5 seconds and half taking about 3 for an average of 4. At the extreme right, all the coins were falling from the right box and taking 4 seconds, and inversely at the left extreme, where all the coins were falling from the left box. For every point along the x -axis, the weighted average was 4 seconds. Given an averaging window of 40 trials, and an average of 4 seconds per trial, the subject could experience the entire reward functions in 160 seconds of coin-dropping time.

What is the rational strategy here? Indifference, which is to say, a 50–50 split, seems indicated, since all strategies earn equal amounts of money. Instead, of the 24 subjects in Experiment 1 and displayed in Figure 1, 22 came within a few percentage points of equalizing the falling speed of

right and left coins, which is at .3 on the x -axis.⁴ Each point shows the average allocation of a single subject during the latter half of the choices made by the subject during the one session he or she ran. The average of all subjects, shown by the crossed lines, was a 30.08 percent choice of the right.

Although the strategy is not rational in any obvious sense, it makes intuitive sense. Picture yourself in the situation and imagine that you find coins from the right box falling more rapidly than coins from the left. Consequently, you start choosing the right side more often. Gradually, coins from the right slow down and coins from the left speed up, and you find that those from the left are now falling more rapidly than those from the right. Your choices will now begin to swing back toward the left, since those coins are falling more rapidly. This strategy leads to an oscillation around 30 percent choice of the right key, and, judging from the data, it is the one adopted by virtually all our subjects. At 30 percent, the subject has equalized the average rate of returns from the two choices, namely a penny per 4 seconds. The equalizing of average rates of return from competing alternatives was first observed experimentally in 1961 in an experiment on pigeons earning bits of food (see my 1961 article). It has since been approximately confirmed in several hundred experiments on various species and is called the matching law because it calls for a match between the ratio of behavioral investments to the yield of those investments across all competing alternatives.

The process of comparing the rates of return and shifting toward the alternative that is currently yielding the better return is called melioration (my 1982 article; myself and Prelec, 1991; myself and Vaughan, 1980). Melioration has the inevitable effect of stabilizing in the vicinity of the matching law, and it also seems to be the process that

⁴Points deviate from the theoretical average line because, with the averaging procedure we used, the current ordinate value for a given proportion of right and left choices depends slightly on the precise sequence of choices within the previous 40 trials.

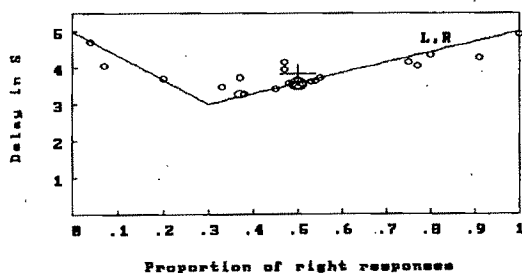


FIGURE 2. EXPERIMENT 2

most people spontaneously invoke when they are asked to describe their own choices.

Here is a second experiment.⁵ Again, volunteer subjects earn coins by striking left or right keys, the coins are worth one cent each, and the time to drop determines how much is earned in a 15-minute session. And again, the time to drop depends on the distribution of choices in the preceding 40 trials, as shown in Figure 2. The x-axis expresses the subject's previous 40 choices as a proportion of choices to the right key. The three lines of Figure 1 are perfectly superimposed in Figure 2. Whatever a subject has done in the preceding 40 choices, coins fall at the same speed from either box.

Suppose, for example, that a particular subject has divided his choices 50-50 and now makes a choice. Whether he chooses left or right, the coin will take about 3.6 seconds to fall. If he had been choosing only the left for the previous 40 trials, a coin from either the left or right would take 5 seconds to fall. The fastest possible speed is 3 seconds, which is what either the right or left coin would take if 30 percent of the preceding 40 choices had been of the right. Since the delays are always equal for right and left coins, the weighted average delay is also equal to them.

The most profitable strategy is at the lowest point on the line, where coins are all taking 3 seconds to fall. But instead of choosing right 30 percent, the 24 subjects are spread out broadly, approximating a 50-50 division between left and right. The

average choice, the crossed line, was 50.08 percent on the right. Many of the individual subjects are also near indifference, as well as their average. Indifference is what the process of melioration implies, for the average yield for left choices is always the same as that for right choices.

Melioration, rather than maximization, appears to have driven the subject's choices in these two experiments. In Experiment 1, melioration defined a particular allocation of choices, that subjects approximated, even though all allocations yielded equal overall earnings. In Experiment 2, maximization called for a particular allocation, but melioration did not; the subjects seemed to be indifferent to the effect of their allocations on earnings. The coin-dropping times involved in the two experiments were essentially the same. The sessions lasted equal times and the amounts of money were the same. In Experiment 1, a 2-second difference was enough to determine behavior; in Experiment 2, a 2-second difference was ignored. From these experiments, we could postulate that a 2-second difference controls behavior if, and only if, it is relevant to melioration.

But neither experiment pitted melioration against maximization in a way that would decisively affect earnings. In the first experiment, melioration caused the subject to obey the matching law when doing so had no effect on money. In the second, indifference to the two alternatives cost the subjects something like a half-second per choice, which may be considered negligible. Let us now consider a third experiment in which melioration implies the maximally *inefficient* allocation of choices.

The procedure is summarized in Figure 3.⁶ As before, the time taken for a coin to drop depends on the subject's choices. The abscissa here is the subject's allocation to the right key during the just-preceding ten choices. The entire reward functions could be experienced in a minute or less. Coins were again worth one cent. Coins from the

⁵Also by Vaughan and myself.

⁶Experiment by Loewenstein, Prelec, Vaughan, and myself.

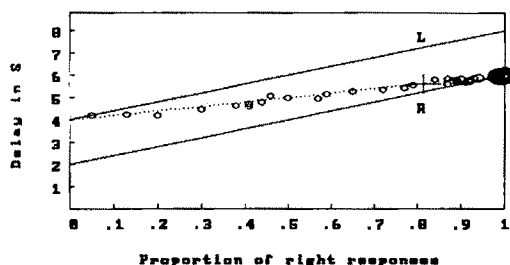


FIGURE 3. EXPERIMENT 3

right box always took 2 seconds less to fall than coins from the left. However, the more choices there were of the right box in the preceding ten trials, the more slowly coins dropped from either box. The optimal strategy would have been to choose the left key on every trial (disregarding the negligible "end effect," which no subject exploited). Melioration dictates choosing right on every trial: it dictates maximal inefficiency. For the 56 subjects, the average chose the right on 82 percent of the trials. A large majority of subjects "learned" the least efficient strategy. At most, one subject discovered the maximizing strategy. Obeying the principle of melioration gets relatively expensive here.

II. Conclusions and Implications

In some of the procedures we have examined, the average performance is closer to maximizing earnings than here presented. For example, when the interaction between choice and consequence is more rapid (when the averaging window is, say, 3 or 6 preceding trials rather than 10 or 40), performance approximates maximization. This suggests an exceedingly limited capacity for taking account of the effect of one's own choices on marginal rates of return. In contrast, our subjects were quite sensitive to average rates of return, as shown by the close approximation to the matching point in Experiment 1.

Another factor appears to be the dimension of reward. Times rates of money, in the guise of a falling coin on a computer monitor, was the relevant dimension here. In other experiments, coin denomination has been the relevant dimension. When the pre-

ceding choices determined what the coins were worth, performances were closer to maximization. It takes a larger averaging window to produce a given degree of suboptimality with coin denomination as the reward dimension than coin rate.

Yet another class of factors affecting performance concerns the degree to which the alternatives are presented as discretely competing responses. The left and right arrow keys were perfectly correlated with coins dropping from left and right boxes, respectively. But when subjects were able to choose a certain proportion of coins from one box or the other, much like a subscription to left or right coins over some interval of time, performance shifted toward maximization. It seems that anything that makes it easier for a subject to redefine the response categories will make it easier to maximize.

For some experimental conditions, subjects span the range between maximization and matching. Individual differences have occasionally looked trimodal—with some subjects near the matching point, some near the maximizing point, and some near 50–50, perhaps indifferent or confused (or both). Intersubject variability seems to peak under conditions that are transitional between those producing maximization and those producing melioration. For example, as the averaging window is increased from, say, 6 to 40, the intersubject variability rises, then falls, as the average subject moves from maximizing earnings to obeying the matching law.

At the beginning of the paper, I suggested that the evidence points toward generically suboptimal human choice under at least some conditions, and possibly under all conditions. A word of explanation about the latter possibility seems in order, inasmuch as the evidence seems only to say that people are sometimes irrational in their choices, not that they are always irrational. But there is a difference between generic suboptimality and universal irrationality. It should be obvious that, for any situation for which a maximization strategy exists, there also exists a definition of response alternatives such that the process of melioration leads to maximization. Given this, it could

be argued that people are, in fact, always following the principle of melioration and that, when they are being rational, they are being rational only incidentally. This argument would be worth making only if there were also a well-developed theory of how a subject defines response alternatives, and we are not yet at that point.

Indeed, we do not know the learning algorithm that produces matching. Melioration implies a comparison of current average yields from competing alternatives, but beyond that vague characterization, the process remains to be explicated. At present, the prospects for doing so are unclear. We have examined the trial-by-trial choice patterns of our subjects and the only conclusion I can draw is that people appear to have many different ways to reach the matching point. Some subjects converge rapidly on an allocation, others scan the range of allocations more or less systematically, and still others do not behave in a way that allows simple characterization. For animal subjects as well, individual variation in the trial-by-trial choices is much greater than the variation around the matching point (see John Bailey and James Mazur, 1990). In the behavioral, as in the physical, sciences, the stable equilibria are often discovered before the dynamic processes that produce them.

REFERENCES

- Bailey, John T. and Mazur, James E., "Choice Behavior in Transition: Development of Preference for the Higher Probability of Reinforcement," *Journal of the Experimental Analysis of Behavior*, May 1990, 53, 409-22.
- Bradshaw, Christopher M. and Szabadi, Elemer, "Quantitative Analysis of Human Operant Behavior," in G. Davey and C. Cullen, eds., *Human Operant Conditioning and Behavior Modification*, New York: Wiley & Sons, 1988, 225-59.
- Davison, Michael and McCarthy, Diane, *The Matching Law: A Research Review*, Hillsdale: Erlbaum, 1988.
- Herrnstein, Richard J., "Relative and Absolute Strength of Response as a Function of Frequency of Reinforcement," *Journal of the Experimental Analysis of Behavior*, July 1961, 4, 267-72.
- _____, "Melioration as Behavioral Dynamism," in M. L. Commons et al., eds., *Matching and Maximizing Accounts*, Cambridge: Ballinger, 1982, 433-58.
- _____, and Prelec, Drazen, "Melioration: A Theory of Distributed Choice," *Journal of Economic Perspectives*, forthcoming 1991.
- _____, _____, and Vaughan, William, Jr., "An Intra-Personal Prisoners' Dilemma," unpublished manuscript, Harvard University, 1986.
- _____, and Vaughan, William, Jr., "Melioration and Behavioral Allocation," in J. E. R. Staddon, ed., *Limits to Action*, New York: Academic, 1980, 143-76.
- Williams, Ben A., "Reinforcement, choice, and response strength," in R. C. Atkinson et al., eds., *Stevens' Handbook of Experimental Psychology*, Vol. 2, New York: Wiley & Sons, 1988, 167-244.

Artificial Adaptive Agents in Economic Theory

By JOHN H. HOLLAND AND JOHN H. MILLER*

Economic analysis has largely avoided questions about the way in which economic agents make choices when confronted by a perpetually novel and evolving world. As a result, there are outstanding questions of great interest to economics in areas ranging from technological innovation to strategic learning in games. This is so, despite the importance of the questions, because standard tools and formal models are ill-tuned for answering such questions. However, recent advances in computer-based modeling techniques, and in the subdiscipline of artificial intelligence called machine learning, offer new possibilities. Artificial adaptive agents (AAA) can be defined and can be tested in a wide variety of artificial worlds that evolve over extended periods of time. The resulting *complex adaptive systems* can be examined both computationally and analytically, offering new ways of experimenting with and theorizing about adaptive economic agents.

Many economic systems can be classified as complex adaptive systems. Such a system is *complex* in a special sense: (i) It consists of a network of interacting agents (processes, elements); (ii) it exhibits a dynamic, aggregate behavior that emerges from the individual activities of the agents; and (iii) its aggregate behavior can be described without a detailed knowledge of the behavior of the individual agents. An agent in such a system is *adaptive* if it satisfies an additional pair of criteria: the actions of the agent in its environment can be assigned a value (performance, utility, payoff, fitness, or the like); and the agent behaves so as to increase this value over time. A complex adaptive system, then, is a complex system

containing adaptive agents, networked so that the environment of each adaptive agent includes other agents in the system.

Complex adaptive systems usually operate far from a global optimum or attractor. Such systems exhibit many levels of aggregation, organization, and interaction, each level having its own time scale and characteristic behavior. Any given level can usually be described in terms of local niches that can be exploited by particular adaptations. The niches are various, so it is rare that any given agent can exploit all of them, as rare as finding a universal competitor in a tropical forest. Moreover, niches are continually created by new adaptations. It is because of this ongoing evolution of the niches, and the perpetual novelty that results, that the system operates far from any global attractor. Improvements are always possible and, indeed, occur regularly. The everexpanding range of technologies and products in an economy, or the everimproving strategies in a game like chess, provide familiar examples. Adaptive systems may settle down temporarily at a local optimum, where performance is good in a comparative sense, but they are usually uninteresting if they remain at that optimum for an extended period.

A theory of complex adaptive systems based on AAA makes possible the development of well-defined, yet flexible, models that exhibit emergent behavior. Such models can capture a wide range of economic phenomena precisely, even though the development of a general mathematical theory of complex adaptive systems is still in its early stages.¹ The AAA models complement current theoretical directions; they are

*Professor of Psychology and Computer Science and Engineering, University of Michigan, Ann Arbor, MI 48109; and Assistant Professor of Economics and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, respectively; both are External Professors, Santa Fe Institute, Santa Fe, NM 87501.

¹It is important in this research to determine just where the potential for general solutions exists. There are simple models of cellular automata, for example, wherein the solutions to particular questions are computationally irreducible—the shortest way to analyze the system is to run the complete computation.

not intended as a substitute. Many of the most interesting questions concern points of overlap between AAA models and classical theory. As a minimal requirement, wherever the new approach overlaps classical theory, it must include verified results of that theory in a way reminiscent of the way in which the formalism of general relativity includes the powerful results of classical physics.

I. Why Study Artificial Adaptive Agents?

The AAA models have several characteristics that are not available in traditional modeling techniques. Models based on pure linguistic descriptions, while infinitely flexible, often fail to be logically consistent. Mathematical models lose flexibility, but gain a consistent structure and general solution techniques. The AAA models, specified in a computer language, retain much of the flexibility of pure linguistic models, while having precision and consistency enforced by the language. The resulting models are dynamic and are "executable" in the sense that the unfolding behavior of the model can be observed step by step. This makes it possible to check the plausibility of the behavior implied by the assumptions of the model. The precision of the definitions also opens AAA models to mathematical analysis. The ability to explore a wide range of phenomena involving learning and adaptation, linked with the rigor imposed by a computer language, provides a powerful modeling technique.²

The AAA models offer a way of approaching one of the major questions of present theory. Current theoretical constructs, based on optimization principles, often require technically demanding derivations. It is an obvious criticism of these constructs that real agents lack the behavioral sophistication necessary to derive the proposed solutions. This dilemma is resolved if it is postulated that adaptive mechanisms, driven by market forces, lead the

agents to act as if they were optimizing (see, for example, Milton Friedman, 1953). AAA explicitly model this link between adaptation and market forces, and can thus be used to analyze the conditions under which optimization behavior will (not) occur.

Insofar as human behavior is driven by adaption, an understanding of AAA may prove to be a useful benchmark for, and provide insights into, existing human experiments (see, for example, J. A. Andreoni and Miller, 1990; Brian Arthur, 1990).³ An experiment consisting of artificial agents allows the utility, risk aversion, information, knowledge, expectations, and learning of each subject to be carefully controlled. Moreover, at any point in the experiment, the knowledge and learning of the artificial agents can be "reset" to any desired previous state, and subtle variations of the environment can be analyzed. The strategy (as well as the behavior) of the AAA can always be explicitly analyzed, something not usually possible with human subjects. Finally, the infinite patience and low motivational needs of AAA "subjects" implies that large-scale experiments can be conducted at a relatively low cost.

A major feature of AAA models is their ability to produce emergent behavior. A wide variety of behaviors can arise endogenously, even though these behaviors, as with any model, are constrained by the initial structure. The possibilities are so rich that it is often difficult to predict on a priori grounds what behaviors and structures will emerge. It thus becomes possible to explore realms that were unanticipated when the model was defined. Analysis of these emergent phenomena should offer both insights and suggestions for new theorems about the effects of adaptive agents in economic systems.

The AAA models may also prove useful in studying economic systems that have either an absence or a plethora of theoretical solutions. Many important economic prob-

²Programming even a simple market is instructive on the limitations of both the pure linguistic and mathematical approaches.

³Artificial agents could also be used as "subjects" in pilot studies to identify potentially interesting new human experiments.

lems, such as double-auction strategies, multisectoral general equilibrium models, and the like, have no easily derived analytic solutions. Several AAA techniques were originally designed as optimization methods for environments that are nonlinear, noisy, discontinuous, or involve enormous search spaces. As a result, they offer useful numerical techniques for such problems in economics. At the opposite extreme are systems with multiple solutions. For example, in repeated games, the Folk theorem often admits a vast number of potential solutions. In these cases, the interaction of the adaptive systems with the economic environment may narrow the set of potential solutions. Different equilibria may have different degrees of *adaptive complexity*.

Beyond complementing current theoretical and empirical work, AAA offer the potential for unique extensions of current theory. The mechanisms generating the global behavior of a complex adaptive system can be directly observed when the computer is an integral part of the theory. For such theories, the computer plays a role similar to the role the microscope plays for biology: It opens up new classes of questions and phenomena for investigation. Problems that prove difficult for traditional mathematical approaches are often easily implemented as an AAA system. In that form, they can be dissected and modified with ease, providing new opportunities for theory generation and testing. More generally, the potential for the development of a general calculus of "adaptive mechanics" exists. A calculus of these systems would combine the advantages of analytic perspicacity with computer-driven hypothesis testing.

II. Some Current Artificial Adaptive Agent Techniques

A wide range of computer-based adaptive algorithms exist for exploring AAA systems, including classifier systems, genetic algorithms, neural networks, and reinforcement learning mechanisms. The multiplicity of techniques presents a problem for analysis. How sensitive are the results to a particular incarnation of the adaptive agent? This

problem, of course, confronts any attempt to lessen the rationality postulates traditionally used in economic theory. Usually, there is only one way to be fully rational, but there are many ways to be less rational. It is important in building a theory based on AAA to construct agents that exhibit robust behavior across algorithmic choices. Current economic studies of adaptive agents rely on genetic algorithms (R. M. Axelrod, 1987; Miller, 1989; Andreoni-Miller) and classifier systems (R. Marimon et al., 1990; Arthur).

Genetic algorithms (GAs) were developed by Holland (1975) as a way of studying adaptation, optimization, and learning. They are modeled on the processes of evolutionary genetics. A basic GA manipulates a set of structures, called a *population*. Structures are usually coded as strings of characters drawn from some finite alphabet (often binary). For example, in a game context, a string might be interpreted either as a simple strategy (a rule table) or as a computer program for playing the game (a finite automaton). Depending upon the model, an agent may be represented by a single string, or it may consist of a set of strings corresponding to a range of potential behaviors. For example, a string that determines an oligopolist's production decision could either represent a single firm operating in a population of other firms, or it could represent one of many possible decision rules for a given firm. Whatever the interpretation, each string is assigned a measure of performance, called its *fitness*, based on the performance of the corresponding structure in its environment. The GA manipulates this population in order to produce a new population that is better adapted to the environment.

In execution, a GA first makes copies of strings in the population in proportion to their observed performance, fitter strings being more likely to produce copies. As a result, fitter strings are more likely to contribute to the new population. After the copies are produced, they are modified by the application of genetic operators. The genetic operators provide for the introduction of new strings (structures) that still

retain some of the characteristics of the fitter strings in the parent population.

The primary genetic operator for a GA is the *crossover* operator. The crossover operator is executed in three steps: 1) a pair of strings is chosen from the set of copies; 2) the strings are placed side by side and a point is randomly chosen somewhere along the length of the strings; 3) the segments to the left of the point are exchanged between the strings. For example, crossover of 111000 and 010101 after the second position produces the offspring strings 011000 and 110101. Crossover, working with reproduction according to performance, turns out to be a powerful way of biasing the system toward certain patterns, *building blocks*, that are consistently associated with above-average performance.

It can be proved (see Holland, 1975) that GAs are a powerful technique for locating improvements in complicated high-dimensional spaces. They exploit the mutual information inherent in the population, rather than simply trying to exploit the best individual in the population. We can liken each of the potential building blocks to one arm of an n -armed bandit. Under this interpretation, each successive generation samples the building blocks in a way that closely corresponds to the optimal solution of an n -armed bandit problem. The GA learns by biasing the search toward combinations of above-average building blocks. Reproduction and crossover are very simple operations that impose low-information and processing requirements on the agents employing them.

A *classifier system* (CS) (Holland et al., 1986) is an adaptive rule-based system that models its environment by activating appropriate clusters of rules. It uses a GA to revise its rules. Each rule is in condition/action form, and many rules can be active simultaneously. The action part of a rule specifies a message that is to be posted when the rule is activated. The condition part of a rule specifies messages that must be present for it to be activated. Thus, each rule is a simple message-processing device that emits a specific message when certain other messages are present. Overt actions affecting the environment are the result of

messages directed to the system's output devices (effectors), while information from the environment is received via messages generated by its input devices (detectors). The overall system is computationally complete in the sense that any program written in a programming language, such as FORTRAN, can also be implemented by a CS.

A CS-rule does not automatically post its message when its condition part is satisfied. Rather, it enters a competition with other rules having satisfied conditions. The outcome of this competition is based on a quantity, called *strength*, assigned to each rule. A rule's strength measures its past usefulness, and it is modified over time by one of the system's learning algorithms (see below). There may be more than one winner of the competition at any given time—hence a cluster of rules can react to external situations. A CS operates on large numbers of rules, with a small number of simple, domain-independent mechanisms. It provides emergent, learned capabilities for reacting to its environment.

A CS adapts or learns through the application of two well-defined machine-learning algorithms. The first algorithm, called a *bucket-brigade algorithm*, adjusts rule strengths. Each rule is treated as an intermediate producer in a complex economy, buying input messages and selling output messages. When a satisfied rule R succeeds in the competition to post its own message, it pays the rule(s) that supplied the messages satisfying its condition part. This amount is subtracted from R 's strength. On the next time-step, if other rules are satisfied by R 's message, and win the competition in turn, then R receives the rules' payment. R 's strength is increased accordingly. The net effect of the two transactions is R 's profit (loss). Some rules also act directly on the environment in a way that produces direct payoff from the environment to the system. Their strength is increased in proportion to that payoff. A rule's strength will increase over time only if it earns a profit, on average, in these transactions. Generally this happens only if the rule directly produces payoff, or else belongs to one or more causal chains leading to payoff. Under appropriate conditions, the

strengths assigned by the bucket-brigade algorithm do converge to a useful measure of the rule's contributions to system performance (Holland et al.).

In order to generate and test new approaches to the environment, the CS needs a second learning algorithm, a *rule discovery algorithm*. A GA can be used for this purpose, because the rules of a CS can be represented by strings in an appropriate alphabet, and a rule's strength amounts to a measure of its performance. The GA, by forming new rules in terms of tested, above-average building blocks, transfers experience from the past to new situations. Plausible new rules result—rules to be tested and retained or discarded on the basis of their ability to enhance the performance of the CS.

Under the combined effects of the bucket-brigade and genetic algorithms, rules become coupled in complex networks. Clusters and hierarchies of rules emerge. Over time, these substructures serve as building blocks for still more complex substructures. A CS agent can: 1) generate broad categories for describing its environment (so that experience can be brought to bear on novel situations); 2) progressively refine and elaborate the relation between categories (using experience to make distinctions and associations not previously possible); 3) use these categories to build internal models that supply the agent with expectations about the world; 4) treat all internal models as provisional (subject to confirmation or refutation as experience accumulates); and 5) generate new hypotheses that are plausible in terms of accumulated experience. Moreover, because of the bucket-brigade algorithm, these activities can proceed in an environment where payoff is intermittent or rare. Such capacities enable a CS agent that is not omniscient to act with increasing rationality.

III. Towards a Mathematics of Complex Adaptive Systems

A mathematical calculus appropriate to the study of complex adaptive systems must meet distinctive requirements. The usual mathematical tools, exploiting linearity,

fixed points, and convergence, provide only an entering wedge. In addition we need a mathematics that works in close conjunction with computer modeling techniques—one that puts more emphasis on combinatorics and algorithms. We require techniques that emphasize the emergence of structure, particularly internal models, through the generation, combination, and interaction of building blocks. The present situation seems quite similar to that of evolutionary theory prior to the development of a mathematical theory of genetic selection (R. A. Fisher, 1930).

Though there is nothing like an overall theory, there are some extant pieces of mathematics that are relevant. The schema theorems for genetic algorithms (Holland, 1975) offer some insight into processes that discover and recombine building blocks. It appears that schema theorems are special cases of a much more general formulation of the effects of recombination in evolution. This formulation should bring some of the more sophisticated tools of mathematical genetics to bear on adaptive agent models. Mathematical work aimed at understanding the evolution of CS may also be useful. The progressive development of hierarchical organization can be treated as the addition of levels to a quasi homomorphism (Holland et al.).

Perpetual novelty can be modeled by a regular Markov process in which each of the states has a recurrence time that is large with respect to any feasible observation time. Equivalence classes can be imposed and used as the states of a *derived* Markov process (Holland, 1986). Work by Miller and S. Forrest (1989), based on S. A. Kauffman's (1984) studies of random graphs, provides additional insights into the emergent structures of CSs.

IV. Conclusions

The AAA research complements ongoing theoretical and empirical work, allowing exploration and analysis of previously inaccessible phenomena. What are the future prospects for this line of inquiry? Early work with AAA in economics has shown that they can acquire sophisticated behavioral patterns. Observation of the course of learning

in these AAA has already increased our understanding of some economic issues. Even limited AAA open up new avenues for analyzing decentralized, adaptive, and emergent systems. Steady advances in computation and AAA modeling offer ever more powerful tools for programming artificial worlds. By executing these models on a computer we gain a double advantage: (i) An experimental format allowing free exploration of system dynamics, with complete control of all conditions; and (ii) an opportunity to check the various unfolding behaviors for plausibility, a kind of "reality check." Whether or not agents in such worlds behave in an optimal manner, the very act of contemplating such systems will lead to important questions and answers.

REFERENCES

- Andreoni, J. A. and Miller, J. H., "Auctions with Adaptive Artificially Intelligent Agents," Santa Fe Institute Working Paper, No. 90-01-004, 1990.
- Arthur, W. B., "A Learning Algorithm that Replicates Human Learning," Santa Fe Institute Working Paper, No. 90-026, 1990.
- Axelrod, R. M., "The Evolution of Strategies in the Iterated Prisoner's Dilemma," in L. D. Davis, ed., *Genetic Algorithms and Simulated Annealing*, Los Altos: Morgan-Kaufmann, 1987, 32-41.
- Fisher, R. A., *The Genetical Theory of Natural Selection*, Oxford: Clarendon Press, 1930.
- Friedman, M., *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.
- Holland, J. H., *Adaptation in Natural and Artificial Systems*, Ann Arbor: University of Michigan Press, 1975.
- _____, "A Mathematical Framework for Studying Learning in Classifier Systems," in J. D. Farmer et al., eds. *Evolution, Games and Learning*, Amsterdam: North-Holland, 1986, 307-17.
- _____, et al., *Induction: Processes of Inference, Learning, and Discovery*, Cambridge: MIT Press, 1986.
- Kauffman, S. A., "Emergent Properties in Randomly Complex Automata," *Physica D*, 1984, 10, 145-56.
- Marimon, R., McGratten, E. and Sargent, T. J., "Money as a Medium of Exchange in an Economy with Artificially Intelligent Agents," *Journal of Economic Dynamics and Control*, May 1990, 14, 329-73.
- Miller, J. H., "The Coevolution of Automata in the Repeated Prisoner's Dilemma," Santa Fe Institute Working Paper, No. 89-003, 1989.
- _____, and Forrest, S., "The Dynamical Behavior of Classifier Systems," in J. D. Schaffer, ed., *Proceedings of the Third International Conference on Genetic Algorithms*. San Mateo: Morgan-Kaufmann, 1989, 304-10.

AMERICAN ECONOMIC ASSOCIATION

**PROCEEDINGS
OF THE
HUNDRED AND THIRD
ANNUAL
MEETING**

**WASHINGTON, D.C.
DECEMBER 28-30, 1990**

Minutes of the Annual Meeting Washington, D.C. December 29, 1990

The one hundred and third annual meeting of the American Economic Association was called to order by President Gerard Debreu at 5:55 P.M., December 29, 1990, in the Ballroom of the Sheraton Washington Hotel, Washington, D.C. He began the meeting by reading the items of business on the agenda and the order of business.

The first item to be considered was the minutes of the last annual meeting as published in the *Papers and Proceedings* issue of the *American Economic Review* (May 1990, pp. 453–54). It was moved and seconded to approve the minutes as published. The motion passed without discussion.

The next items considered were the reports of the Secretary (C. Elton Hinshaw), Treasurer (Hinshaw), Editor of the *American Economic Review* (Orley Ashenfelter), Editor of the *Journal of Economic Literature* (John Pencavel), Co-editor of the *Journal of Economic Perspectives* (Carl Shapiro) and Director of *Job Openings for Economists* (Hinshaw). Each discussed his written report (published elsewhere in this issue) which was available to the members prior to the meeting. After the Secretary's report, James Angel asked him what was the probable impact on hotel room rates of moving the annual meeting dates from December 28–30 to the first week in January. Hinshaw responded that there probably would be some increase because a few more conventions and trade shows met during January—competition for hotel space was somewhat heavier. The December period was a notoriously slack time for the hotel industry. However, the Secretary did inform the assembly that single room rates in New Orleans (1992) would be \$60 and in Anaheim (1993), \$63. Carl Lundgren questioned the Secretary about the availability of the agenda prior to the business meeting. He wondered if the Secretary deliberately made it difficult to obtain a copy in order to discourage attendance. Hinshaw responded

that the agenda was made available during the recess between the Presidential Address and the business meeting, and at the AEA booth in the registration area from the beginning of the annual meetings on December 27th. The Executive Committee would be delighted if more members attended the business meeting. (The Secretary notes that just prior to the President's address, the Chair always reminds the audience that the business meeting will follow the address, informs them that the agenda will be available during the recess period, and encourages them to stay for the meeting. The Chair repeats this message again immediately after the address and prior to the recess.) Lundgren stated that no one at the registration site was able to tell him where copies of the business agenda might be obtained prior to the presidential address and that the business agenda was not made available at the business meeting site until after nearly all the attendees at the presidential address had already left.

Ashenfelter reported that the *AER* would probably move to double-blind refereeing—the author does not know who the referee is, the referee is not informed who the author is. He had been testing for the impact of double-blind refereeing on the editorial process at the *Review* by randomly assigning part of the submissions to a double-blind refereeing process. The experiment is completed and a report of the results will be available in 1991. Evidence indicated that the rejection rate of papers that went through the double-blind process was significantly higher than that of the single-blind process (referee knows who the author is but not vice versa). Kenneth Arrow asked if the *AER* had not used double-blind refereeing in the past. Ashenfelter responded that George Borts had used it, but that Robert Clower had abandoned it. Ashenfelter stated that the *AER*'s Board of Editors had voted to adopt such a process.

It was widely perceived to be a fairer system.

Brian Horrigan inquired about publishing one or more of the journals in an electronic format. Pencavel responded that the bibliographic section of the *JEL* was now available in electronic form, on-line through DI-ALOG and on compact disc (CD-ROM) from SilverPlatter Information, Inc. He was not aware of any plans to extend electronic publishing to other AEA journals.

Following the completion of the above listed reports, the Chair called for other business to come before the assembly. Carl Lundgren moved the following resolution, which had been submitted in writing prior to the meeting:

*Resolution to Democratize the
Selection of A.E.A. Officers*

The current wording of Article IV, Section 2, of the Bylaws of the American Economic Association shall be deleted and amended to read as follows:

"Section 2. At least 60 days before the annual meeting the President-elect of the Association shall appoint a Nominating Committee, this Committee to consist of a past officer as non-voting Chairman and not less than seven other randomly selected members of the Association. Members of the Nominating Committee must agree to be present at the meeting of the Nominating Committee, which shall take place one day prior to the main activities of the annual convention, and in the same general place. The President-elect shall send out a sufficient number of invitations to members of the Association, each member having an equal chance of being randomly selected for the invitation, to ensure that at least seven members of the Association agree to serve on the Nominating Committee.

The Nominating Committee shall present to the Executive Committee one nomination for President-elect, two nominations for Vice-President, and two nominations for elected member of the Executive Committee. The Nominating Committee shall also have power to recommend removal or cen-

sure of any elected or appointed officer or employee of the Association. After the Nominating Committee has presented its nominations, the Executive Committee shall decide upon one additional nomination for President-elect, two additional nominations for Vice-President, and two additional nominations for elected member of the Executive Committee.

The nominations of elected officers by the two Committees shall be presented to the members at the annual meeting for discussion and nonbinding vote. If the members present are dissatisfied with the nominees presented, they may by majority vote choose to consider additional nominees. If they so choose to consider additional nominees, they shall be entitled to place on the mail ballots up to one additional nominee for President-elect, up to two additional nominees for Vice-President, and up to two additional nominees for elected member of the Executive Committee.

The actual election of officers by the membership shall take place by mail ballot conducted by the Secretary each year. The ballot shall list all candidates alphabetically, with indications of how many votes each candidate received in the nonbinding vote and which of the two Committees (or the annual meeting) made the nomination. The Secretary shall mail the ballots as soon as practical within six weeks after the annual meeting. The Secretary shall also set a deadline for receipt of ballots, not less than four weeks after the mailing and not more than twelve weeks after the annual meeting. The candidates with the highest number of votes for the various offices shall be elected. The results of the election shall be certified and announced by the Secretary in the *American Economic Review* and at the next annual meeting.

During each mail balloting, the members shall be solicited concerning their suggestions for future nominees to the offices of the Association. The Secretary shall prepare for the Nominating Committee and the Executive Committee a list of all the names suggested, and the number of members making the suggestion, omitting only

such names as are not members of the Association. Also for their use, the Secretary shall solicit from each member whose nomination was suggested by any member, issue and resume information concerning his/her possible candidacy, except that if more than thirty persons have been suggested for any one office, the Secretary may limit these solicitations to not less than thirty persons who have been suggested by the largest number of members."

The Executive Committee is hereby directed by the members assembled to approve the above amendment and to submit it by mail ballot to the members at large.

The Chair asked if anyone in the assembly wished to have Lundgren read the resolution. No one did. The Chair asked if there was a second to the motion. There was not. Lundgren stated that the resolution had already been seconded in writing by Edwin G. Olson, thirty days in advance of the meeting. The Chair ruled that the resolution could only be seconded by someone who was present at the meeting. The motion died on the floor for lack of a second.

The Chair again called for other business. He recognized Lundgren, who moved the following resolution which had been submitted in writing prior to the meeting:

*Resolution to Require Discussion of
Amendment to the Bylaws*

Article VI of the Bylaws which currently reads, "Amendments, after having been approved by a majority of the Executive Committee, may be adopted by a majority of votes cast in a mail ballot," shall be deleted in its entirety.

Article V of the Bylaws shall be amended to insert the following sentences: "Proposed amendments which have been approved by a majority of the members at an annual meeting may be adopted by a majority of votes cast in a mail ballot. Proposed amendments which have been approved by a majority of the Executive Committee may be adopted by a majority of votes cast in a mail ballot, provided that a

resolution concerning the amendment was discussed and voted on by the members at either of the two preceding annual meetings."

The above sentences shall be inserted just before the last sentence of Article V which currently reads: "Mail ballots shall be accompanied by (1) a brief statement by the sponsor or sponsors in support of the resolution, (2) a brief statement of the views of the Executive Committee and, if the Committee favors or takes no position on the resolution, (3) a brief statement by a person or persons whom the Committee designates to represent opposing views."

The Executive Committee is hereby directed by the members assembled to approve the above amendment and to submit it by mail ballot to the members at large.

The Chair asked if anyone wished to have the resolution read. No one so indicated. The Chair asked if there was a second to the motion. There was no second from the floor. The motion died.

The Chair asked if there was other business to come before the assembly. He recognized Lundgren, who moved the following resolution which had been submitted in writing prior to the meeting:

*Resolution to Rescind an Unauthorized
Resolution of the March 1974
Executive Committee Meeting*

The resolution of the March 8, 1974 meeting of the Executive Committee (*AER*, May 1975, p. 449), which purports to impose restrictions on what resolutions may be considered by the members at their annual meeting, is hereby rescinded. This resolution is rescinded because (a) the Executive Committee is not authorized by either the Charter or the Bylaws to impose restrictions on what resolutions the members may properly consider, and (b) a general resolution of this nature is ill-advised, since it fails to allow for legitimate exceptions to the proposed restrictions.

It is recommended for good order that members who know at least 60

days in advance that they would like to bring a resolution before the annual meeting should submit their resolution in writing to the Secretary at least 30 days in advance of the annual meeting. It is recommended (but not required) that the annual meetings refuse to consider surprise resolutions that could have been written up at least 60 days in advance of the meeting.

If the Secretary receives a copy of any proposed resolution which has been proposed and seconded by members of the Association, both being members in good standing, the Secretary shall immediately acknowledge receipt of the resolution and shall immediately submit copies to both the President and Counsel. Within 14 days of receiving the text of the resolution, both the President and the Counsel shall state in writing their opinions concerning whether the resolution should be considered in order or out of order and, if out of order, the reasons why. If the resolution is stated to be out of order, the sponsors shall be invited to present counter-arguments before the Executive Committee, if the sponsor(s) so choose.

If a written resolution has been properly proposed and seconded and submitted to the Secretary at least 30 days in advance of the annual meeting, the Secretary shall reproduce the proposed resolution and make copies available at least one hour in advance of the annual business meeting. Additionally, the Secretary shall do at least one of the following: (a) Distribute copies of the proposed resolution to the membership upon registration at the annual convention, or (b) post copies of the resolution in conspicuous locations prior to registration and leave posted until the annual business meeting is adjourned.

The Chair inquired if anyone wished to have the resolution read. John Roemer asked that he be given one minute to read the resolution for himself. The Chair granted one minute of reading time to the assembly. After the minute had passed, the Chair asked if there was a second to

Lundgren's motion. Roemer seconded it. Lundgren was again granted the floor. He favored the motion for reasons given in the resolution and expanded upon them. He noted that there is no *quid pro quo* in the present procedure. Members have to propose motions 30 days prior to the meeting, but the officers of the Association can rule them out of order without prior notice to the proposer of the motion. An "out of order" ruling without prior notification is unfair. It deprives one of the freedom of speech. It indicates that the person is not worth listening to. Requiring prior notice of such a ruling seemed reasonable. He recalled that a previous resolution of his had been ruled out of order by the Executive Committee in a meeting to which he was not allowed to attend. He thought that to be unfair. He further stated that the Secretary had not acknowledged receipt of his resolutions until December 18th, indicating that he did not think that soon enough. He ended by indicating that sufficient information about the agenda was not made available and this explained the poor attendance at the business meeting. He expressed some doubt that the resolutions were available when and where the Secretary had earlier in the meeting stated them to be. The Secretary apologized to Lundgren for the late acknowledgement of his resolutions. He stated that it was his practice to inform first those who had submitted resolutions that were not in good order. He apparently had implicitly been assuming that proposers of resolutions would assume that "no news was good news," that is, the resolutions had been received. The Secretary added, however, that a resolution was not necessary to get him to send an acknowledgment. He would mend his ways. Roemer asked about the restrictions imposed on resolutions that were referred to in Lundgren's resolution. The Secretary responded that the restriction was a time restriction—resolutions had to be received by the Secretary 30 days prior to the annual meeting so that they could be duplicated in time to distribute at the meetings. Robert Eisner added that not all resolutions, even if submitted properly, would necessarily be in order. Some resolu-

tions were outside the stated purposes of the Association and in conflict with the charter and bylaws. Such resolutions would be ruled "out of order" by the Chair. The President, not the Executive Committee, makes such a ruling. He, Eisner, had been President when Lundgren submitted his first resolution. As Chair of the meeting, he had ruled it out of order. (The interested reader is referred to the minutes of the 1988 annual business meeting, *Papers and Proceed-*

ings issue of the *American Economic Review*, May 1989, p. 391.) The Chair then called for the vote. The motion failed with only one person voting in favor.

There being no other business before the assembly, Debreu introduced the incoming President, Thomas Schelling, who adjourned the meeting at 6:30 P.M.

Respectfully submitted,
C. ELTON HINSHAW, *Secretary*

Minutes of the Executive Committee Meetings

Minutes of the Meeting of the Executive Committee in Washington, D.C., March 23, 1990.

The first meeting of the 1990 Executive Committee was called to order at 10:00 A.M., March 23, 1990, in the Congressional Room of the Sheraton Washington Hotel in Washington, D.C. Members present were Gerard Debreu (presiding), George Akerlof, Orley Ashenfelter, Gregory Chow, Rudiger Dornbusch, Robert Eisner, Stanley Fischer, C. Elton Hinshaw, Allan Meltzer, John Pencavel, Isabel Sawhill, Thomas Schelling, and Joseph Stiglitz. Present as Counsel was Leo Raskind. Present for a part of the meeting were members of the 1990 Nominating Committee: Gary Becker, Duncan Foley, Edward Lazear, Frank Levy, John Shoven, and Janet Yellen.

Minutes. The minutes of the meeting of December 27, 1989, were approved as written and circulated prior to the March meeting.

Report of the Secretary (Hinshaw). The 1989 meeting in Atlanta attracted 6,627 registrants, a record number for a meeting away from the East Coast. The previous high was 6,384 at the 1987 Chicago meeting. Forty-one other associations, societies, and organizations met with us, 428 scholarly sessions were organized, and 114 events (lunches, cocktail parties, committee meetings, etc.) scheduled.

The Secretary reported that he had investigated the possibility of meeting in San Francisco and Chicago in 1993. Both cities reported conflicts with already booked conventions. The largest hotel in Los Angeles is owned by Portman Properties. This eliminates L.A. as a potential site for the time being. Both Seattle and San Diego would require the use of a convention center and shuttle services, things we try to avoid. The Secretary noted that the growth in attendance was beginning to present problems for site selection. Fewer cities had the hotel facilities required for the meetings. Most cities would now require the use of a convention center or shuttling or both. The discussion indicated that, in general, we should continue to try to avoid both conven-

tion centers and busing in city selection. It was VOTED to authorize the Secretary to negotiate contracts with the hotels in Anaheim, California for the 1993 meeting.

Since the option of *not* receiving one of the journals with a \$6 reduction in dues has become available, some 7,583 persons have renewed their memberships or joined. Ninety-one percent (91%) elected to receive all three journals; 2% elected not to receive the *American Economic Review*; 3%, the *Journal of Economic Literature*; and 3.5% the *Journal of Economic Perspectives*. Five people seem to have elected not to receive any of the three.

After a discussion of the proposal submitted by Trans National Financial Services, it was VOTED not to establish an affinity credit card program. Concern was expressed that the Association would be viewed as endorsing a specific credit card by entering into such a contract. The Association did not wish to do that.

It was VOTED to submit an amendment to Article V, Section 2 of the bylaws to a vote of the members. The proposed amended bylaw would read as follows:

A resolution adopted at an annual meeting in which less than five percent of the membership of the Association has voted thereon shall be submitted to a vote by mail ballot no later than the ballot for officers if a majority of the Executive Committee determines that, because of the nature or consequences of the resolution, all members should have the opportunity to participate in the final decision.

Dan Newlon, of the National Science Foundation, had inquired if the Association had an interest in developing an annual "directory" of *all* academic economists, not just AEA members. The NSF would find such a directory of great value. Newlon had generated a couple of proposals from others, which had been sent to the Secretary, to develop such a data base, and he would like the AEA's participation. The existing proposals were not sufficiently well-defined

to reach a judgment, but it was decided that the subject should be pursued by the Secretary.

Report of the Editor of the American Economic Review (Ashenfelter). It was VOTED to approve the Editor's recommendation of reappointment of the following persons to the Board of Editors: George Akerlof, Jo Anna Gray, John F. Kennan, and Richard Roll. With the same vote, his recommendation of John Campbell, Robert Hodrick, and Hal Varian as new members was also approved. In response to a question, Ashenfelter stated that the size of the Board had grown in number, primarily because each of the co-editors needed members in their specific areas. Board members were used heavily as referees. He also reported that the report on double-blind refereeing might be finished this summer; he expected to give the results at the next meeting.

Report of the Editor of the Journal of Economic Literature (Pencavel). Pencavel reported that the Association had entered into a contract with SilverPlatter to put the bibliographic section of the *JEL* on a compact disk (CD-ROM). The *JEL* will provide the data and help market the disk; SilverPlatter will develop the software, produce the disks, and also market the product. The new product does not replace the on-line data base with DIALOG; it simply makes the same data base available in a different electronic form. There was some concern that the pricing structure effectively eliminated the individual user. It was suggested that it might be feasible to make the disk available with a year's lag at a price that individuals could afford. He further reported that he hoped to have the new classification scheme in place by this summer. It is currently being reviewed for final revisions.

Report of the Editor of the Journal of Economic Perspectives (Stiglitz). Stiglitz reported that he was in the process of arranging symposia on several topics, some of which are Third-World Debt, Bubbles, The Role of the State in Development, Intellectual Property Rights, and The Economic Status of Blacks.

Report of the Director of Job Openings of Economists (Hinshaw). Hinshaw reported

that he had received a letter from a member, Carl Lundgren, proposing a change in the subscription policy to *JOE*. Currently only AEA members may subscribe. Lundgren proposes that nonmembers be allowed to subscribe at a rate that is not higher than average cost. The Director agreed with the proposal and recommended that a new category of subscriber—Institutions and Nonmembers—be established and that the rate be \$25 (plus postage if outside the United States). It was VOTED to accept the Director's recommendation.

The 1990 Program (Schelling). Schelling reported that currently he had organized 51 sessions for the meetings, 22 of which are scheduled to be published in the *Papers and Proceedings* issue of the *AER*. Titles of some of the sessions are Post-Communist Economic Transition, Rumania and Bulgaria, Hungary and Poland, Czechoslovakia, Yugoslavia, and East German, China, East and West Europe in 1992, Tax Reform, Learning, Space, The Nonrational in Decisions, Greenhouse Warming, Altruism, Illicit Drugs, and Conflict.

After Schelling completed his description of the program for the coming year, the discussion turned to the issue of "openness" of the program. There was some suggestion that perhaps the selection of papers to be published in the *Papers and Proceedings* might be made more open if papers had to be submitted prior to the meetings and without the "guarantee" of publication. If publication was not automatic, better papers may result. Some thought that the distinctive nature of the *Papers and Proceedings*, the short deadlines involved, and the prerogatives of the Presidents to have the program reflect their interest made "competitive" publication less feasible and less desirable. No action was taken.

Report of the Nominating Committee (Becker). Becker, who chaired the Committee, reported the following nominations for the indicated offices: for Vice-President—Henry J. Aaron, Claudia D. Goldin, Robert E. Hall, and Daniel McFadden; for the Executive Committee—Michael J. Piore, T. N. Srinivasan, Martin L. Weitzman, and Gavin Wright. The Nominating Committee and the Executive Committee, acting together as an

electoral college, nominated William Vickrey as President-elect and elected Victor Fuchs and Merton Miller as Distinguished Fellows.

Report of the Treasurer (Hinshaw). Audited financial statements for 1989 were distributed prior to the meeting. The statements show an operating deficit of \$450 thousand, up from \$398 thousand in 1988. After taking into account investment income recognized of \$297 thousand, the overall deficit was \$153 thousand. The operating loss was \$21 thousand smaller than budgeted and investment income was \$32 thousand larger than the \$265 thousand budgeted. The overall deficit was \$53 thousand less than anticipated. Cash and investments increased from \$5.9 million in 1988 to \$6.4 million at the end of 1989.

The Budget Committee had met that morning and made the following general policy recommendation: Increase membership dues annually at the same rate as the increase in the Consumer Price Index until such time as the ratio of net worth to annual expenditures equals one. This policy presupposes that such increases will continue to result in operating deficits and that such deficits should be tolerated until the Net Worth/Annual Expenditures ratio decreases to one. If the ratio rises much above where it now is, say to two, the policy should be reconsidered. This could happen if there was a significant increase in the value of the Association's portfolio of investments.

In keeping with this general policy, the Budget Committee recommended that 1991 dues be increased by the increase in the CPI from June 1989 to June 1990. It was VOTED to increase dues by the recommended amount. It was understood that the increase would be rounded to the nearest dime and that the Treasurer could, if necessary because of time constraints for printing renewal notices, etc., change the dates for calculating the change in the CPI.

Other Business. It was VOTED to reappoint Ashenfelter to a third term as editor of the *American Economic Review*. It was VOTED to reappoint Pencavel to a third term as editor of the *Journal of Economic Literature*. Both terms would end December

31, 1994. There followed a discussion of the desirability of establishing a more systematic procedure for reviewing the performance of the editors and the quality of the journals. Both editors thought such a procedure would be useful. Although there seemed to be a consensus that the suggestion was a good one, no specific action was taken.

An *Ad Hoc* Committee of three (Meltzer, Akerlof, and Dornbusch) was appointed to arrange three or four sessions for the 1990 meetings devoted to discussing (teaching) some of the more recent and useful research techniques that have been developed, a kind of "continuing education" section of the AEA program. It was noted that there is a substantial lag between the time new innovations are developed and their appearance in textbooks. These sessions would substantially reduce that lag. They might best be described as lectures on chapters that might appear in a graduate textbook with an accompanying reading list.

The final item of discussion was techniques and strategies for increasing membership in the Association. The Secretary noted that he had instituted an exchange of ads with two Japanese journals but had taken no further action. Several possibilities were put forward—ads in the *Economist*, the *New York Times*, and the *Wall St. Journal*, buying various mailing lists and doing a direct mailing, etc. Ashenfelter volunteered to investigate the cost of obtaining the mailing list of the *Economist*. The Secretary will inquire of other social science societies what procedures they use to seek new members.

The meeting was adjourned.

Minutes of the Meeting of the Executive Committee in Washington, D.C., December 27, 1990.

The second meeting of the 1990 Executive Committee was called to order at 10:00 A.M., December 27, 1990, in the Congressional Room of the Sheraton Washington Hotel in Washington, D.C. Members present were Gerard Debreu (presiding), George Akerlof, Orley Ashenfelter, Gregory Chow, Rudiger Dornbusch, Robert Eisner, Stanley Fischer, C. Elton Hinshaw, Allan Meltzer, John Pencavel, Susan Rose-

Ackerman, Thomas Schelling, and Lawrence Summers. Present as guests were Henry Aaron, Claudia Goldin, Michael Piore, Carl Shapiro, and William Vickrey. Present for the purpose of giving reports were Nancy Gordon, Lee Hansen, Anne Krueger, and Margaret Simms. Present as Counsel was Leo Raskind.

Debreu opened the meeting by greeting the guests and reviewing the agenda. The first item of business was the approval of the minutes of the March 23, 1990, meeting. These were approved as written.

Report of the Secretary (Hinshaw). The Secretary added his welcome to the new members of the 1991 Executive Committee and thanked those of the present Committee whose terms were expiring—Akerlof, Dornbusch, Eisner, Meltzer, and Sawhill. He then informed the group that the next annual meeting would be held in New Orleans on January 3–5, 1992, and the 1993 meeting in Anaheim, January 5–7. The Secretary said that cities currently under consideration for 1994 were Boston, Denver, and New York. He, once again, reviewed some of the problems being generated by the growth in the number of sessions, events, and attendance. This growth, combined with the reluctance to use convention centers, caused him to recommend that the meetings be extended to three full days. Sessions can no longer be accommodated within a two and a half day framework. After discussing other options such as having another period for sessions that would begin about 4:00 or 4:30 P.M., it was agreed that the meetings would be extended to include the afternoon of the third day. He reported that the “telephone directory” would be published during the fourth quarter of 1991. It would contain name, full address (the one to which the journals are sent), office phone and fax numbers, but no biographical information. It will be sent to members free of charge. It was VOTED to authorize a charge of up to but no more than \$10.00 to non-members and subscribers for the telephone book.

Since October of last year, members have had the option of not receiving one of the three journals and having their dues re-

duced by \$6.00. Since then some 21,000 persons have renewed their memberships or joined the Association. Of these, some 1,985 elected not to receive one of the journals. Four hundred seventy-one (471) elected not to receive the *AER*; 716, the *JEL*; and 787, the *JEP*. Eleven people have chosen to receive none of the three.

The Secretary had recently received an order for mailing labels of the membership for the purpose of sending a fund-raising appeal for University Scholarships for South African Students, Inc., which claims to be a nonprofit, tax-exempt organization. He declined to sell the membership list for this purpose. He is currently authorized to sell the list for commercial purposes. As far as he could determine, the list had never been sold for the purpose of soliciting contributions to worthy causes. The list is currently sold primarily to publishers. If a request is received that is not a “standard” one, the Secretary makes an *ad hoc* decision based on his judgment of the appeal of the proposed mailing to the membership. He asked for advice or instruction about what the policy should be. It was VOTED to authorize the Secretary to sell the membership list to all who requested it for a period of two years. The experiment would then be evaluated to determine if a change in the policy was needed. The Secretary understood that some modicum of inquiry might be undertaken to determine the “legitimacy” of the potential purchaser.

Report of the Editor of the American Economic Review (Ashenfelter). Ashenfelter briefly reviewed his written report which is published elsewhere in this issue and circulated to the Committee prior to the meeting. It was VOTED to approve his recommendation to reappoint James E. Anderson, appoint Kyle W. Bagwell, D. John Roberts, and Suzanne A. Scotchmer, and extend by one year the term of David Sappington as members of the *AER* Board of Editors.

He also reported that the experiment on double-blind refereeing which had been underway for about two years was over and that the final report would be published sometime in 1991. The one definite result to

be reported now was that double-blind refereeing makes a statistically significant difference in the acceptance rate for publication—the double-blind acceptance rate is only about 70 percent of the single-blind rate. As a result of this finding, Ashenfelter thought that the *AER* would move to the double-blind refereeing process. The Board of Editors was scheduled to discuss the issue later at these meetings.

Report of the Editor of the Journal of Economic Literature (Pencavel). Pencavel highlighted some of the items presented in his written report which was circulated prior to the meeting and is published elsewhere in this issue. It was VOTED to approve his recommendation to appoint Sherman Robinson to a three-year term on the Board of Editors. He announced that the new classification scheme which had been under consideration for the past two years would be introduced during 1991. Further, the Economic Literature Index is now available in the form of compact discs through Silver-Platter Information, Inc. as well as on-line through DIALOG. Initial sales of the disc are encouraging. Currently under review are the criteria for selecting journals to be included in the bibliographic sections of the *JEL*.

Report of the Editor of the Journal of Economic Perspectives (Shapiro). In the absence of Joe Stiglitz, Co-editor Shapiro reviewed the written report which is published elsewhere in this issue. Two new features have been added: "Retrospectives" and "Policy Watch." These replace, in part, "Puzzles" and "Anomalies." Some of the symposia to be published in 1991 are "Intellectual Property," "Trade Liberalization," and "Socialist Economies." It was VOTED to approve the recommendation to appoint Peter Murrell and Gene Grossman to the Board of Editors for terms of three years.

Report of the Director of Job Openings for Economists (Hinshaw). Hinshaw reviewed his written report which is published elsewhere in this issue. He then called the Committee's attention to a special situation that had arisen during the past year. An AEA member, Pinaki Sankar Bose, had

written him to complain about the hiring procedures of Concordia University (Montreal, Canada). The barest facts seem to be that Bose was offered in writing a three-year, tenure-track position by the Chair of the Economics Department. Bose accepted and notified his other employment options that he had done so. Subsequently, the Dean of Concordia unilaterally changed the official offer to a one-year, nontenure track appointment for reasons that did not entail the qualifications of Bose. Bose wrote the Secretary to complain. The Secretary wrote (May 17, 1990) the Rector of Concordia, stating the allegations of Bose and asking the Rector to review the facts, judge the validity of the allegations, and ascertain what responsibility Concordia had to Bose.

In October, Concordia submitted an ad for *JOE*. As Director of *JOE*, Hinshaw wrote (October 19, 1990) the Chair of the Economics Department informing him that he had received no response from the University concerning Bose's allegations and informed him that if Concordia wished to use the services of the Association to help recruit faculty members, the Association needed assurance that its help would be used appropriately and responsibly. Hinshaw said he would not publish the ad without a response to his inquiry of May 17th.

The Rector responded in a letter dated October 24, 1990, in which, in part, he gave his assurances that such an incident would not occur again and steps had been taken to prevent such an event in the future. Hinshaw's reading of his response led him to conclude that (1) Bose's allegations concerning his dealings with Concordia University were accepted as true by the University, (2) Bose was the victim of an internal misunderstanding within the University, (3) no official action was taken by the University from the time of Hinshaw's first inquiry on May 17th to his second of October 19th, and (4) the Rector had concluded to do nothing to remedy the damage done Bose except send him a formal letter of apology, dated October 24, 1990. Based on his understanding of all that had transpired, Hinshaw decided to deny publication of Con-

cordia's ad in the December issue of *JOE* and bring the situation to the Executive Committee's attention to determine what it might wish to do.

It was VOTED to affirm the Director's decision to deny Concordia University the privilege of advertising in the December issue of *JOE* and to extend the prohibition to August 1, 1991.

The Committee then discussed if a "warning" to job applicants should be published in *JOE* to alert them to the sometimes subtle distinction made by some universities between "informal" offers by Chairs and "formal" offers by Deans since the Association takes very seriously the responsibility of universities to make offers in good faith, to be careful about who makes commitments for it, and to live by commitments made on its behalf. No action was taken.

Report of the Committee on the Status of Women in the Economics Profession (Gordon). Gordon commented on her written report which had been circulated to the Executive Committee prior to the meeting and is published elsewhere in this issue. The discussion that followed concerned the slow progress of women through the professorial ranks. Despite the substantial growth of women in the profession during the past decade, this increase is not appearing as yet in the senior ranks. The proportion of assistant and associate professors who were women approximately tripled between 1974 and 1989, rising from 8 to 20 percent and from 3 to 9 percent respectively; the proportion of full professors who were women grew only from about 2 to 3 percent.

Report of the 1991 AEA Program Chair (Vickrey). Vickrey, President-elect for 1991, reported that proposals for sessions and abstracts of papers had begun to flow into his office. He had not yet settled on a focus for the program.

Report on the Status of Minorities in the Economics Profession (Simms). Simms is the new Chair of the Committee, having taken over from Ron Oaxaca. She reported that the Committee had recommended that the Summer Program be awarded to Stanford

University for a period of three years with an option to renew for two more. Tom Macurdy will be the Director. The Committee currently has funds from the Ford Foundation sufficient to operate the program for another three years. A written report is published elsewhere in this issue. It was suggested that the Committee consider giving higher stipends to fewer participants.

Report of the Commission on Graduate Education (Krueger and Hansen). Krueger reported that the Commission had finished its work and would present its recommendations at a session at the current meetings. The final report along with a background paper by Hansen would be available. Current discussions were underway with the Editors of the AEA journals about which journal would be the most appropriate place to publish the Commission's findings. Commercial publishers were available to publish the Commission's work in monograph form. The Commission recommended that the AEA hold the copyright and receive any royalties. It was VOTED to accept any royalties. It was understood that Hansen, in consultation with Raskind, would negotiate the contract with prospective publishers. It was also understood that the Treasurer would keep track of the amount of royalties received from this source as a potential reminder that a follow-up study might be undertaken in five or so years. It was VOTED to thank and commend the Commission for its work.

Other Business. It was VOTED to commend the *Ad Hoc* Committee (Meltzer, Akerlof, and Dornbusch) that arranged the "continuing education" lectures for these meetings and to reappoint them and ask them to do it again for the 1992 meetings.

During the course of the discussion of three resolutions by Carl Lundgren and one by Suresh Deman, it was VOTED to endorse President Debreu's decision not to invite Lundgren to attend the meeting of the Executive Committee. It was also VOTED to submit to the membership an amendment to the bylaws that would make resolutions passed at the annual AEA business meeting with fewer than half of one

percent of the membership present and voting advisory only. This latter motion was passed following a long discussion lamenting the poor attendance at the business meetings. With some regularity, resolutions are passed with fewer than 50 members voting. Suggestions were offered about how to increase attendance (for example, change the time). The Secretary reminded the group that the business meeting was in its current time slot because attendance had been worse when it was held separately from the Presidential Address and when the Presidential Address had been held at 8:00 P.M. It was also suggested that more signatures be required on resolutions before they could be considered. During a discussion of the nominating process, considerable support was expressed for some review, examination, and possible changes in the procedures. It was understood that the Executive Committee wished to reexamine the nominating process, perhaps by a committee ap-

pointed for that purpose. Ultimately, the only actions taken were those reported above.

Finally, the Committee considered a suggestion from a member concerning the establishment of a new award to honor that economist or group of economists who had contributed most to the analysis and discussion of public policy issues. The member had suggested the award be named after Joe Pechman. The Committee, although apparently sympathetic with the idea, deferred any action.

Report of the Treasurer (Hinshaw). Hinshaw reviewed his report which is published elsewhere in this issue. It was VOTED to approve the 1991 budget as submitted.

There being no further business, the meeting adjourned at 5:50 P.M.

Respectfully submitted,
C. ELTON HINSHAW, *Secretary*

Report of the Secretary for 1990

Telephone Directory. The American Economic Association will publish a "telephone directory" of members in late 1991. It will be generated from the journal mailing list and will include the name, address, and telephone/fax numbers of AEA members who have active memberships as of June 1991. A space is now provided on both the membership application and renewal notice for correcting telephone numbers and adding a fax number.

Annual Meetings. The next annual meeting will be held in New Orleans, January 3-5, 1992. Placement Service will open for business one day earlier (January 2) than the meetings. There is no meeting in 1991.

Elections. In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee and the Electoral College.

The Nominating Committee, consisting of Gary S. Becker, Chair, Duncan K. Foley, Edward P. Lazear, Frank S. Levy, John B. Shoven, Peter Temin, and Janet L. Yellen submitted the nominations for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and Executive Committee meeting together, selected the nominee for President-elect. No petitions were received nominating additional candidates.

President-elect
William Vickrey

<i>Vice-President</i>	<i>Executive Committee</i>
Henry J. Aaron	Michael J. Piore
Claudia D. Goldin	T. N. Srinivasan
Robert E. Hall	Martin L. Weitzman
Daniel McFadden	Gavin Wright

The Secretary prepared biographical sketches of the candidates and distributed ballots last summer. On the basis of the canvass of ballots, I certify that the following persons have been duly elected to the

respective offices:

President-elect (for a term of one year)
William Vickrey
Vice-Presidents (for a term of one year)
Henry J. Aaron
Claudia D. Goldin
Executive Committee (for a term of three years)
Michael J. Piore
Gavin Wright

In addition, I have the following information:

Number of legal ballots	5,034
Number of invalid envelopes	244
Number of envelopes received after October 1	29
Number of envelopes returned	5,307

The proposed amendment to the bylaws was adopted; the bylaws as amended now read:

Article V, Section 2

2. A resolution adopted at an annual meeting in which less than five percent of the membership of the Association has voted thereon shall be submitted to a vote by mail ballot no later than the ballot for officers if a majority of the Executive Committee determines that, because of the nature or consequences of the resolution, all members should have the opportunity to participate in the final decision.

Membership. The total number of members and subscribers is shown in Table 1.

TABLE 1—MEMBERS AND SUBSCRIBERS
(End of Year)

	1988	1989	1990
Class of Membership			
Annual	17,410	18,172	17,863
Junior	1,939	2,170	2,533
Life	353	338	318
Honorary	33	33	33
Family	437	470	446
Complimentary	475	387	385
Total Members	20,647	21,570	21,578
Subscribers	5,793	5,736	5,785
Total Members and Subscribers	26,440	27,306	27,363

The total has fluctuated between 25,000 and 27,500 since 1975. This year it reached an all-time high of 27,363, an increase of 57 over last year.

National Registry. The National Registry for Economists continues to be operated on a year-round basis by the Illinois State Employment Service. Economists looking for jobs and employers are urged to register. This is a placement service that maintains the anonymity of employers. The Association is indebted to the Registry for assistance and supervision at the employment service provided at the annual meetings. Employers are reminded of the Association's bimonthly publication, *Job Openings for Economists*, and their professional obligation to list their job openings.

Permission to Reprint and Translate. Official permission to quote from, reprint, or translate and reprint articles from the *American Economic Review*, *Journal of Economic Literature* and the *Journal of Economic Perspectives* totaled 602 in 1990, compared to 696 in 1989. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to obtain the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of \$150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

Staff. The staff in the Secretary's office is unusually loyal, gracious, and efficient. I am indebted to them as is the Association. They are Mary Winer, Kim Adair, Norma Ayres, Dana Coleman, Sherry Davis, Marlene Hall, Dana Jackson, and Violet Sikes.

Committees and Representatives. Listed below are those who served the Association during 1990 as members of committees or representatives. The year in parentheses indicates the final year of the term to which they were appointed. On behalf of the Association, I thank them for all their service.

AEA COMMITTEES 1990

Budget Committee

C. Elton Hinshaw, Chair

Isabel V. Sawhill (1990)
Lawrence H. Summers (1991)
Gregory C. Chow (1992)
Gerard Debreu (1990)
Thomas C. Shelling (1991)

Census Advisory Committee

Margaret C. Simms, Chair (1991)
Ben E. Laden (1990)
Dale T. Mortensen (1990)
Victor Zarnowitz (1990)
Timothy F. Bresnahan (1991)
Michael McKelvey (1991)
Thomas W. Synnott III (1991)
Gail Fosler (1992)
Bronwyn H. Hall (1992)

Committee on Economic Education

John Siegfried, Chair (1990)
Michael J. Boskin (1990)
Robert J. Highsmith (1990)
Robert T. Michaels (1990)
Robin L. Bartlett (1991)
Donald N. McCloskey (1991)
Phillip Saunders, Jr. (1991)
William J. Baumol (1992)
Michael K. Salemi (1992)
William E. Becker, Jr., *ex officio*

Economics Institute Policy and Advisory Board

Edwin S. Mills, Chair (1990)
Samuel A. Morley (1990)
Koichi Hamada (1991)
Ray Marshall (1991)
Daniel W. Bromley (1992)
James A. Millar (1992)
John P. Evans (1993)
Lawrence J. Lau (1993)

Finance Committee

C. Elton Hinshaw, Chair
Robert Hamada (1990)
B. Douglas Bernheim (1991)
Richard S. Peterson (1992)

Commission on Graduate Education in Economics

Anne O. Kreuger, Chair
W. Lee Hansen, Executive Director
Kenneth J. Arrow
Olivier J. Blanchard
Alan S. Blinder

Claudia Goldin
Edward E. Leamer
Robert J. Lucas
John C. Panzar
Rudolph G. Penner
T. Paul Schultz
Joseph E. Stiglitz
Lawrence H. Summers

Committee and Honors and Awards

Oliver E. Williamson, Chair (1991)
Dale W. Jorgenson (1991)
Alan S. Blinder (1993)
Hollis Chenery (1993)
John B. Taylor (1993)
Paul A. David (1995)
Andreu Mas-Colell (1995)

1990 Nominating Committee

Gary S. Becker, Chair
Duncan K. Foley
Edward P. Lazear
Frank S. Levy
John B. Shoven
Peter Temin
Janet L. Yellen

Committee on Political Discrimination

Burton A. Weisbrod, Chair (1992)
Benjamin J. Cohen (1990)
Clark W. Reynolds (1990)
Stephen Marglin (1992)
Wassily Leontief (1992)

*Committee on the Status of Minority Groups
in the Economics Profession*

Margaret C. Simms, Chair (1993)
William A. Darity, Jr. (1990)
William D. Bradford (1990)
Susan M. Collins (1990)
George J. Borjas (1991)

Vernon J. Dixon (1991)
Clifford E. Reid (1991)
Ronald L. Oaxaca (1990)

*Committee on the Status of Women in the
Economics Profession*

Nancy M. Gordon, Chair (1991)
Shulamit Kahn (1990)
Kathryn Morrison (1990)
Barbara Newell (1990)
Elizabeth Hoffman (1991)
Shelly J. Lundberg (1991)
June O'Neill (1991)
Daniel H. Newlon (1991)
Rebecca Blank (1992)
Marjorie H. Honig (1992)
Barbara Wolfe (1992)
Myrna H. Wooders (1992)
Gerard Debreu, *ex officio*
Joan G. Haworth, Membership Secretary

Committee on U.S.-China Exchanges

Gregory C. Chow, Chair
Kenneth Arrow
Richard H. Holton
Lawrence R. Klein

Committee on U.S.-Soviet Exchanges

Lawrence R. Klein, Chair (1990)
Zvi Griliches (1990)
Edward A. Hewett (1990)
Kenneth Arrow (1991)
Robert Eisner (1991)
Franklyn D. Holzman, (1991)
Leonid Hurwicz (1991)
Stanley Johnson (1991)
Richard Quandt (1991)
Nathan Rosenberg (1991)
Richard N. Rosett (1991)
Gerard Debreu, *ex officio*

COUNCIL AND OTHER REPRESENTATIVES

AAAS Consortium of Affiliates for International Programs

C. Elton Hinshaw (1992)

*American Association for the Advancement
of Science, Sec. K, Social Economics and
Political Sciences*

Adam Rose (1992)

American Council of Learned Societies

C. Elton Hinshaw (1990)

*Consortium of Social Science Associations
(COSSA)*

Isabel V. Sawhill (1992)
C. Elton Hinshaw

Council of Professional Associations on Federal Statistics (COPAFS)

Charles R. Hulten (1991)

Edward F. Denison (1992)

International Economic Association

Kenneth Arrow (1990)

Policy Board of the Journal of Consumer Research

Louis L. Wilde (1991)

National Bureau of Economic Research

David A. Kendrick (1990)

Social Science Research Council

Robert M. Coen (1990)

REPRESENTATIVES OF THE ASSOCIATION ON VARIOUS OCCASIONS—1990

Inaugurations

Julius Wesley Becton, Jr.,

Prairie View A&M University

Sherry S. Zwiebel

Robert I. Rotberg, Lafayette
College

Marietta Constantinides

Daniel L. Ritchie, University
of Denver

Kishore G. Kulkarni

John Michael Palms, Georgia
State University

Romie Tribble

Roger Harold Hull, Union College

Elena H. L. Alvarez

John Thomas Casteen III,
University of Virginia

Judy Klein

Thomas H. Kean, Drew University

Elizabeth C. Bogan

Lattie F. Coor, Arizona State
University

Taeho Kim

C. ELTON HINSHAW, *Secretary*

Report of the Treasurer for the Year Ending December 31, 1990

The proposed budget for 1991 in Table 1 projects an operating loss of \$637 thousand, investment income of \$252 thousand, and an overall deficit of \$385 thousand. The deficits result primarily from the decision to begin publication of the *Journal of Economic Perspectives* without an increase in real dues. Net worth at the beginning of 1991 is expected to exceed budgeted expenditures by more than a million dollars so drastic action is not required. Another deficit can be accommodated.

Audited statements for 1990 will be published in the June issue of the *American Economic Review*.

I wish to thank Norma Ayres, our accountant, and Mary Winer, the Administrative Director, for their valuable help and patience in assisting me in carrying out the duties of the Treasurer.

C. ELTON HINSHAW, *Treasurer*

TABLE 1—1991 BUDGET, AMERICAN ECONOMIC ASSOCIATION
(Thousands of dollars)

	First Nine Months (Unaudited)		Actual	Full Year Budgeted	
	1989	1990	1989	1990	1991
REVENUES FROM DUES AND ACTIVITIES					
Membership dues	\$772	\$826	\$1,049	\$1,041	\$1,113
Nonmember subscriptions	560	572	759	757	775
Subtotal	1,332	1,398	1,808	1,798	1,888
Subscriptions, <i>Job Openings for Economists</i>	25	25	38	34	40
Advertising	101	110	134	140	149
Sale of <i>Index of Economic Articles</i>	100	181	91	197	212
Sales of copies, republications, handbooks	35	37	40	45	45
Sale of mailing list	41	46	57	60	70
Annual meeting	83	82	83	50	50
Sundry	63	79	83	100	85
Total Operating Revenue	1,780	1,958	2,334	2,424	2,539
PUBLICATION EXPENSES					
<i>American Economic Review</i>	586	636	744	849	900
<i>Journal of Economic Literature</i>	702	764	951	1,065	1,085
<i>Journal of Economic Perspectives</i>	292	331	393	428	463
<i>Survey of New Members</i>	52	52	70	70	50
<i>Job Openings for Economists</i>	39	39	62	59	65
<i>Index of Economic Articles</i>	40	205	50	100	68
Subtotal	1,711	2,027	2,270	2,571	2,631
OPERATING AND ADMINISTRATIVE EXPENSES					
General and Administrative	290	330	390	410	426
Committees	40	42	64	60	60
Support of other organizations	44	45	59	58	59
Subtotal	374	417	513	528	545
Total Expenses	2,085	2,444	2,783	3,099	3,176
OPERATING GAIN (LOSS)	(305)	(486)	(449)	(675)	(637)
INVESTMENT GAIN (LOSS)	223	240	297	296	252
SURPLUS (DEFICIT)	(82)	(246)	(152)	(379)	(385)

Report of the Finance Committee

The Finance Committee of the American Economic Association met at the Chicago Club, Chicago, Illinois, at 11:45 a.m. on December 13, 1990. Present were Richard Peterson, Robert Hamada (members of the Committee), and C. Elton Hinshaw (Chairman of the Committee and Secretary-Treasurer of the Association); Robert McNeill and Harvey Hirschhorn (representing Stein Roe & Farnham, investment counsel for the Association).

In 1987, the Committee reviewed recommendations presented by the AEA Committee on Indexing Association Funds concerning the long-term allocation of the Association's investment assets. As a result of that recommendation and the subsequent deliberation of the Finance Committee, it was agreed that the Association's portfolio be comprised of a combination of the Wells Fargo South Africa Free Index Fund, Stein Roe & Farnham's specialty equity mutual funds, and a bond portion managed by Stein Roe & Farnham.

This restructuring took place at the end of June 1988. The current portfolio includes holdings in the Wells Fargo South Africa Restricted Equity Index B, as well as the International Growth, Special and Stock Funds, which are all managed by Stein Roe & Farnham. The Fixed Income portion of the portfolio is currently invested in short/medium-term U.S. government agency notes

and SRF's intermediate-term taxable funds (Managed Bonds & Governments Plus).

As a result of the aforementioned asset allocation restructuring, the overall performance of the Association's fund now reflects the combined efforts of Wells Fargo and Stein Roe & Farnham.

With respect to the 1990 performance (through December 31) of the Association's portfolio, the total return of the account including cash, bonds, and SRF equity funds was -2.6 percent. Focusing on a longer-term period, over the six years ending with December 1990, the total return for the account was +88.3 percent. In dollars, the total return has been approximately \$2,500,000 over the last six years.

After considering the economic outlook and the composition of the Association's portfolio, the Committee approved a 50-70 percent allocation to equities (including the Wells Fargo Fund). In addition, it was decided that international equity exposure should range from 5 to 15 percent, and the minimum cash equivalent position should be 5 percent. The benchmark for performance is a portfolio consisting of 60 percent equities and 5 percent cash equivalents.

Members can obtain a list of the assets in the portfolio by writing the Treasurer.

C. ELTON HINSHAW, *Chair*

Report of the Editor

American Economic Review

The editorial process has continued to work very smoothly during the past year despite some changes in personnel at the *Review*. Our current submission rate (of nearly 1,000 papers per year) has remained fairly stable at a level we first achieved in the early 1980's.

Editorial Process

As Table 1 indicates, the number of submissions has decreased by about 4 percent over last year. Nevertheless, the chances that a submitted paper will eventually be published have decreased slightly from last year to this year. The fraction of submitted papers that we publish is now at an all time historical low of .11.

Table 2 indicates that we published more articles and fewer shorter papers in the *Review* in 1990 than in 1989, and that the total number of pages published dropped slightly. I and my co-editors have adopted a conscious policy of attempting to increase the number of major substantive articles we publish at the expense of shorter papers, comments, and replies. We had some success at doing so during the last year.

Tables 3 and 4, when compared with the results for last year, indicate there has been a small slowdown in the speed with which we handle manuscripts that are ultimately

rejected. A major cause of the delay in handling manuscripts can be attributed, as I indicated in my Report last year, to the use of referees and to the slowness with which we receive reports. From 1989 to 1990, I

TABLE 1—MANUSCRIPTS SUBMITTED
AND PUBLISHED, 1971–90^a

Year	Submitted	Published	Ratio
			Published-to-Submitted
1971	813	115	.14
1972	714	143	.20
1973	758	111	.15
1974	723	125	.17
1975	742	112	.15
1976	695	117	.17
1977	690	114	.17
1978	649	108	.17
1979	719	119	.17
1980	641	127	.20
1981	784	115	.15
1982	820	120	.15
1983	932	129	.14
1984	921	138	.15
1985	952	128	.13
1986	987	123	.125
1987	843	99	.12
1988	844	100	.12
1989	946	116	.12
1990	911	100	.11

^aThe submissions reported for every year refer to the last two months of the previous year and the first ten months of the year reported.

TABLE 2—SUMMARY OF CONTENTS, 1989 AND 1990

	1989		1990	
	Number	Pages	Number	Pages
Articles	51	851	56	932
Shorter Papers, including Comments and Replies	65	427	44	329
Announcements and Notes				
Section		26		28
Index		10		11
Total		1314		1300

and my co-editors have again increased our use of the refereeing process. The result has been an increase in the delay with which we handle manuscripts.

Table 5 indicates that there has been some change from last year in the speed with which accepted papers were published in the *Review*. We continued during 1989-90 to build up a backlog of accepted papers, and this naturally led to an increase in the lag between acceptance and publication. Building up our backlog has, on the whole, been desirable, since we had been operating with far too tight a printer's deadline in the last three years. I now feel that our backlog has stabilized and I suspect that this may lead to a shortening of the time between acceptance and publication of articles in the *Review*.

TABLE 3—DISPOSITION OF MANUSCRIPTS,
1989 AND 1990

	July 1, 1988– June 30, 1989	July 1, 1989– June 30, 1990
Manuscripts Received	923	916
Completed Processing	568	526
Accepted	38	26
Rejected	530	500
Currently in Process	355	390

The subject matter distribution of papers published in the *Review* in 1989 and 1990 is contained in Table 6. It remains my impression that the distribution of published papers reflects fairly accurately the distribution of papers submitted.

Finally, as I reported earlier in the year, we have been testing for the impact of double-blind refereeing on the editorial process at the *Review* by randomly assigning part of our submissions to a double-blind refereeing process. This experiment is now complete and a report of its results will be available in mid-1991.

Papers and Proceedings

The twelfth volume of the *Papers and Proceedings* to be prepared by the editorial staff of the *Review* appeared in May 1990. For the last two years this task has been very capably handled by Ronald Oaxaca (University of Arizona) and Wilma St. John. I am deeply indebted to both of them for the difficult work under extraordinarily tight deadlines that they have so capably performed.

Co-Editors and Board of Editors

There was one major change in the editorial personnel at the *Review* in the last year.

TABLE 4—DISTRIBUTION OF EDITORIAL DECISION LAGS BETWEEN RECEIPT AND REJECTION,
JULY 1, 1989–JUNE 30, 1990

Weeks to Rejection	Total Number of Manuscripts	Percent	No Outside Referees	1 Referee	2 Referees	3 or more Referees
0-4	31	.04	23	6	2	0
5-6	34	.05	1	25	7	1
7-8	54	.07	1	30	21	2
9-10	62	.08	1	27	32	2
11-12	76	.10	0	21	52	3
13-14	73	.10	1	19	50	3
15-16	59	.08	0	13	45	1
17-21	141	.19	0	34	96	11
22-26	73	.10	0	15	55	3
27-30	31	.04	0	7	23	1
31-35	29	.04	0	5	23	1
36-52	48	.07	0	11	31	6
52+	28	.04	0	5	11	12
	739	100.00	27	218	448	46

TABLE 5—AVERAGE PUBLICATION LAGS
BY JOURNAL ISSUE

Journal Issue	Number of Weeks Lag		
	Receipt to Acceptance	Acceptance to Publication	Receipt to Publication
March 1990	58	35	93
June 1990	57	43	100
September 1990	46	47	93
December 1990	59	49	107

Claire Comiskey, our production editor, resigned due to ill health. I am deeply grateful for all the help and effort she provided over the three-year period she served as production editor.

David Baldwin, formerly affiliated with the biology journal *Evolution*, has become our new production editor. His knowledge of scientific writing has already proved invaluable and will, I am sure, help us to improve the clarity of scientific discourse in the *Review*.

I edit the *Review* with the assistance of Robert Haveman (University of Wisconsin), Bennett McCallum (Carnegie Mellon University), and Paul Milgrom (Stanford University). I am deeply indebted to them for the conscientious effort they have expended over the last year.

The Board of Editors now consists of twenty-eight members and I am indebted to

TABLE 6—SUBJECT MATTER DISTRIBUTION OF
PUBLISHED MANUSCRIPTS, 1989 AND 1990

	Published	
	1989	1990
General Economics and General Equilibrium Theory	19	3
Microeconomic Theory	6	21
Macroeconomic Theory	1	5
Welfare Theory and Social Choice	2	1
Economic History, History of Thought, Methodology	1	1
Economic Systems	0	2
Economic Growth, Development, Planning, Fluctuations	10	7
Economic Statistics and Quantitative Methods	5	3
Monetary and Financial Theory and Institutions	4	6
Fiscal Policy and Public Finance	14	10
International Economics	14	9
Administration, Business Finance	4	0
Industrial Organization	12	17
Agriculture, Natural Resources	1	4
Manpower, Labor Population	13	7
Welfare Programs, Consumer Economics, Urban and Regional Economics	10	4
Total	116	100

them all for their efforts. Board members are selected to reflect the highest level of scholarship in the economics profession from the breadth of different fields represented in our submissions. More than fine scholarship is expected of a Board member, however. Board members are also selected because of their conscientiousness, good

TABLE 7—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING, 1990 AER

	Copies Printed	Pages		Cost		
		Net	Gross	Issue	Reprints	Total
March	28,000	312	360	\$67,446.83	\$1,911.50	\$69,358
May	28,000	500	536	95,151.09	3,905.31	99,056
June	28,000	337	360	68,000.98	1,733.10	69,734
September	28,000	347	400	77,654.76	1,533.14	79,188
December ^a	28,000	304	352	68,000.00	1,185.00	69,185
Annual Misc. ^b						12,000
Total		1,800	2,008			\$398,521

^aEstimated.

^bEstimated: based on costs of preparing mailing list, extra shipping charges, and storage costs of back issues.

judgment, and professional reliability. When possible, we like to select Board members from those economists who have been especially helpful in the outside refereeing process.

Two members of the Board completed their terms during 1990: Maurice Obstfeld and Kenneth Singleton. I am most grateful to them and to the continuing members: George Akerlof, James Anderson, Timothy Bresnahan, John Campbell, Henry Farber, Marjorie Flavin, Robert Flood, Claudia Goldin, Jo Anna Gray, Reuben Gronau, Daniel Hamermesh, Robert Hodrick, Kevin Hoover, Kenneth Judd, John Kagel, John Kennan, Edgar Olsen, John Riley, Richard Roll, Thomas Romer, David Sappington, Robert Smith, Barbara Spencer, Richard Tresch, Hal Varian, Kenneth West, John

Wilson, and Leslie Young. I am also grateful for the assistance of many other associates who make it possible to edit and produce the *Review*. I am indebted to our office manager, Shirley Griesbaum, and our editorial assistants, Sandra Grant and Kathy Simkanich, for the fine work they have continued to perform over the past year. I would also like to thank the co-editors' secretaries: Joan Sample (Robert Haveman's office); Pat Niber (Bennett McCallum's office); Kim Mooney and Deborah Johnston (Paul Milgrom's office).

As always, the published version of this Report contains the list of referees who have volunteered their services during 1989. We extend our deepest appreciation for the time and energy they have devoted to the advancement of our science.

A. B. Abel	A. J. Auerbach	E. Berglas	M. D. Bordo
K. G. Abraham	K. M. Ausubel	T. C. Bergstrom	S. Borenstein
D. Abreu	D. K. Backus	J. A. Berkovec	E. R. Borensztein
A. Admati	R. Bade	D. Berkowitz	G. J. Borjas
J. Aizenman	K. Bagwell	B. S. Bernanke	P. Bossaerts
G. A. Akerlof	M. N. Bailly	E. R. Berndt	J. M. Boughton
A. Alesina	S. H. Baker	B. D. Bernheim	J. Bound
L. S. Alexander	N. S. Balke	G. Bertola	A. L. Bovenberg
B. Allen	L. Ball	D. Besanko	M. Boyer
D. W. Allen	C. Ballard	D. Betson	R. S. Boyer
F. Allen	E. Baltensperger	J. Bhagwati	R. Braeutigam
S. D. Allen	W. A. Barnett	J. S. Bhandari	J. A. Brander
S. G. Allen	R. J. Barro	M. Bils	W. H. Branson
J. Alm	T. J. Bartik	J. H. Bishop	R. A. Braun
D. Altig	Y. Barzel	R. Bishop	G. Brennan
J. G. Altonji	K. Basu	D. Bizer	M. Brennan
T. Amemiya	B. W. Bateman	D. A. Black	T. F. Bresnahan
J. E. Anderson	R. G. Batina	C. Blackorby	W. A. Brock
R. W. Anderson	R. C. Battalio	R. M. Blank	S. G. Bronars
A. K. Ando	J. R. Baumgardner	M. Blaug	C. C. Brown
R. Andreano	K. M. Bawn	M. I. Blejer	E. K. Browning
J. Andreoni	M. Baxter	A. S. Blinder	J. K. Brueckner
E. Antoniadou	T. A. Bayoumi	G. Blomquist	J. Bryant
L. Argote	C. M. Beach	R. Blundell	T. Buchmueller
R. J. Arnott	P. J. Beck	R. W. Boadway	W. H. Buiter
D. J. Aron	G. Becker	N. E. Bockstael	J. Bulow
L. Arvan	J. R. Behrman	M. Boldrin	J. Buongiorno
M. Arvin	D. Benjamin	P. Bolton	E. Burmeister
D. A. Aschauer	J. S. Bennett	R. Bolton	L. Buron
W. Asher	B. L. Benson	W. A. Bomberger	G. T. Burtless
A. Atkeson	G. J. Benston	G. Bonanno	J. L. Butkiewicz
D. B. Audretsch	M. C. Berger	E. W. Bond	W. Butz

L. M. Cabral	V. P. Crawford	N. Economides	R. T. Froyen
G. G. Cain	J. Cuddington	H. J. Edison	D. Fullerton
C. W. Calomiris	A. Cukierman	A. Edlin	M. A. Fuss
R. Calvert	W. Cullison	S. Edwards	J. E. Gagnon
G. Calvo	D. M. Cutler	R. G. Ehrenberg	E. Gal-Or
C. F. Camerer	C. d'Aspremont	M. S. Eichenbaum	I. L. Gale
J. Y. Campbell	P. Danzon	R. Eisner	N. Gallini
M. B. Canzoneri	M. R. Darby	R. C. Ellickson	J. Gans
A. S. Caplin	A. F. Daughety	W. Enders	A. M. Garber
D. E. Card	M. David	J. Enelow	B. F. Gardner
J. B. Carlson	C. Davidson	S. L. Engerman	M. R. Garfinkel
D. W. Carlton	S. J. Davis	D. Epple	M. Gavin
J. Carmichael	W. F. De Bondt	N. R. Ericsson	W. T. Gavin
L. H. Carmichael	F. de Leeuw	T. J. Espenshade	M. Gersovitz
M. R. Carter	A. de Palma	C. Evans	A. R. Ghosh
A. C. Case	A. V. Deardorff	G. W. Evans	R. S. Gibbons
S. G. Cecchetti	A. S. Deaton	P. D. Evans	R. J. Gilbert
Y. Chan	E. Dekel-Tabak	P. Eze	I. Gilboa
W. W. Chang	J. B. DeLong	R. C. Fair	C. L. Gilroy
D. Chaplin	P. M. DeMarzo	R. L. Faith	A. Glazer
H. W. Chappell	H. Demsetz	G. Fallis	J. Glazer
V. V. Chari	W. den Haan	R. E. Falvey	J. H. Goddeeris
J. Chavas	J. N. Dertouzos	E. F. Fama	V. P. Goldberg
L. K. Cheng	P. Desai	H. S. Farber	C. D. Goldin
M. Chernew	W. G. Dewald	P. G. Farnham	N. Z. Golding
S. Chew	D. W. Diamond	G. Feder	S. Goldman
P. Chinloy	P. A. Diamond	R. C. Feenstra	M. Goodfriend
R. S. Chirinko	B. T. Diba	C. Fershtman	R. H. Gordon
G. C. Chow	W. T. Dickens	G. S. Fields	R. J. Gordon
C. F. Christ	F. X. Diebold	T. W. Fields	P. Gottschalk
L. J. Christiano	J. DiNardo	S. Figlewski	D. A. Graham
C. C. Chu	E. Dinopoulos	C. C. Findlay	E. Gramlich
P. B. Clark	A. K. Dixit	J. M. Finger	J. A. Gravelle
P. K. Clark	K. M. Dominguez	M. Finn	P. E. Graves
H. R. Clarke	J. B. Donaldson	A. C. Fisher	J. A. Gray
S. Coate	R. G. Donaldson	F. M. Fisher	W. H. Greene
J. H. Cochrane	M. P. Dooley	M. J. Flannery	A. Greif
B. J. Cohen	G. K. Dow	M. A. Flavin	T. A. Gresik
D. Cohen	S. Dowrick	C. J. Flinn	J. M. Griffin
D. C. Colander	D. Dranove	R. P. Flood	Z. Griliches
D. Collie	A. Drazen	R. Forsythe	E. R. Grilli
P. J. Cook	J. Driffill	A. Fraga	V. Grilli
T. Q. Cook	C. Driver	M. Z. Frank	R. Gronau
T. F. Cooley	L. Dudley	R. H. Frank	D. Gros
R. J. Cooper	J. A. Dunlevy	J. A. Frankel	G. M. Grossman
R. W. Cooper	S. Durlauf	B. Fraumeni	H. I. Grossman
C. Corrado	G. P. Dwyer	A. M. Freeman	M. Grossman
S. R. Cosslett	P. H. Dybvig	S. Freeman	H. Grubert
P. N. Courant	R. A. Dye	J. Fried	P. Guidotti
D. Cox	B. C. Eaton	B. M. Friedman	F. Gul
J. C. Cox	J. Eaton	J. W. Friedman	D. D. Haddock
P. C. Cramton	Z. Eckstein	K. A. Froot	G. Hadfield

A. Haight	S. Huddart	R. A. Kerin	R. D. Lee
J. J. Hallman	J. R. Huddleston	J. R. Kesselman	T. K. Lee
J. C. Haltiwanger	G. W. Huffman	M. S. Khan	K. Leffler
D. S. Hamermesh	H. Huizinga	N. M. Kiefer	D. E. Leigh
B. W. Hamilton	J. Huizinga	R. E. Kihlstrom	J. S. Leonard
J. D. Hamilton	C. R. Hulten	J. King	L. L. Leslie
T. H. Hannan	T. Humphrey	M. King	D. Levin
W. L. Hansen	W. Humphries	R. G. King	R. C. Levin
G. A. Hardouvelis	M. D. Hurd	A. Kleidon	D. Levine
J. Harkness	A. M. Husain	B. Klein	J. A. Levinsohn
S. Harrington	S. Husted	P. D. Klemperer	J. Li
M. Harris	R. M. Hutchens	P. Klenow	E. Lichtenberg
G. W. Harrison	A. Hynes	J. L. Knetsch	F. R. Lichtenberg
O. D. Hart	J. Ingersoll	A. Kochar	D. M. Lilien
M. Hashimoto	M. D. Intriligator	L. Kochin	L. A. Lillard
K. Hassett	R. M. Isaac	K. Koford	Y. J. Lin
J. Haubrich	P. Isard	J. E. Kohlhasse	R. C. Lind
T. M. Havrilesky	Y. Ishii	T. E. Kollintzas	P. H. Lindert
F. Hayashi	T. Ito	C. D. Kolstad	C. R. Link
R. W. Haynes	M. Jackson	R. C. Kormendi	P. Linneman
J. J. Heckman	G. C. Jacobson	L. J. Kotlikoff	B. Lipman
R. Heiner	M. Jacobson	D. Kovenock	P. Lipton
W. Heller	R. Jacobson	S. Krasa	S. Lohmann
J. F. Helliwell	A. Jacquemin	R. Krelove	J. Londregan
E. Helpman	A. B. Jaffe	J. J. Kremers	G. Loomes
P. Hendershott	J. A. James	D. Kreps	M. Lundahl
D. W. Henderson	D. T. Jamison	A. O. Krueger	N. Lutz
J. V. Henderson	P. Jefferson	P. R. Krugman	A. B. Lyon
K. Hendricks	B. Jensen	P. Krusell	R. K. Lyons
M. G. Herander	S. Jones	P. J. Kuhn	A. Ma
Z. Hercowitz	P. L. Joskow	H. Kunreuther	L. J. Maccini
B. E. Hermalin	P. Jost	A. Kuprianov	M. Machina
R. J. Herrnstein	B. Jovanovic	K. N. Kuttner	R. Mackay
S. Heston	K. L. Judd	E. R. Kwerel	W. B. MacLeod
R. L. Hetzel	R. E. Just	J. E. Kwoka	W. Magat
A. L. Hillman	F. R. Kaen	F. E. Kydland	S. P. Magee
D. Hirshleifer	J. H. Kagel	A. S. Kyle	G. E. Makinen
J. Hirshleifer	J. A. Kahn	J. M. Lacker	J. M. Malcomson
R. J. Hodrick	M. I. Kamien	J. Laffont	N. G. Mankiw
J. P. Hoehn	G. Kaminsky	K. Lahiri	C. Mann
E. Hoffman	K. Kamiya	D. E. Laidler	W. Manning
T. J. Holmes	E. Karni	R. J. LaLonde	C. Manski
C. A. Holt	H. Katz	J. T. Landa	J. R. Markusen
K. D. Hoover	L. F. Katz	W. Landes	M. L. Marlow
H. Hopenhayn	M. L. Katz	N. R. Lardy	J. Marquez
T. M. Horbulyk	J. B. Kau	L. J. Lau	D. Marshall
J. L. Horowitz	W. Keech	D. Laussel	R. E. Martin
I. J. Horstmann	M. C. Keeley	L. Lave	P. R. Masson
C. W. Howe	D. C. Keenan	E. Lawrence	S. E. Masten
P. Howitt	P. J. Kehoe	E. E. Leamer	D. Mathieson
D. A. Hsieh	A. C. Kelley	S. Lebergott	K. Matsuyama
R. G. Hubbard	J. F. Kennan	J. Ledyard	S. Matthews

C. Matutes	S. Nickell	J. Pratt	M. R. Rosenzweig
C. Maxwell	J. Niehans	E. C. Prescott	S. Ross
K. R. Mayer	L. T. Nielsen	A. Protopapadakis	T. W. Ross
R. P. McAfee	R. G. Noll	D. Purvis	J. J. Rotemberg
B. McCall	W. Nordhaus	Y. Qian	A. E. Roth
J. McCallum	G. Norman	D. Quah	M. Rothschild
D. McDonald	D. C. North	R. E. Quandt	N. Roubini
A. M. McGartland	W. H. Oakland	J. P. Quirk	D. L. Rubinfeld
M. G. McGarvey	W. E. Oates	R. Rahi	G. D. Rudebusch
M. C. McGuire	R. L. Oaxaca	J. Raisian	M. Rush
T. McGuire	A. O'Brien	R. Ram	R. R. Russell
R. I. McKinnon	J. R. O'Brien	V. A. Ramey	J. Rust
J. McMillan	M. Obstfeld	M. R. Ransom	G. L. Salamon
R. Mehra	J. N. Ochs	R. H. Rasche	M. K. Salemi
A. H. Meltzer	E. O. Olsen	D. J. Ravenscraft	M. A. Salinger
J. Melvin	J. I. Ondrich	T. G. Rawski	G. Saloner
M. T. Melvin	J. A. Ordovery	D. Ray	S. C. Salop
E. G. Mendoza	S. Oren	A. Razin	L. W. Samuelson
J. Meyer	M. J. Osborne	R. Ready	W. F. Samuelson
M. Meyer	J. Ostroy	S. Rebelo	T. Sandler
N. Miller	S. Ozler	P. Regibeau	A. M. Santomero
D. Mills	M. Pagano	S. J. Reichelstein	D. E. Sappington
L. J. Mirman	T. R. Palfrey	C. E. Reid	T. J. Sargent
J. A. Miron	A. Panagariya	R. J. Reilly	M. Sattinger
F. S. Mishkin	M. Parkin	J. F. Reinganum	W. Scarth
O. S. Mitchell	D. O. Parsons	P. C. Reiss	D. S. Scharfstein
K. Miyagiwa	B. P. Pashigian	R. Repullo	B. Schefold
H. Miyazaki	D. Patinkin	J. D. Richardson	T. C. Schelling
D. M. Modest	J. Paul	J. Riley	F. M. Scherer
R. A. Moffitt	C. H. Paxson	J. G. Riley	D. E. Schlagenhauf
J. Montgomery	S. Peltzman	M. H. Riordan	F. Schneider
P. J. Montiel	J. Pemberton	V. Rios-Rull	J. K. Scholz
S. Moorthy	C. Pendergast	L. Rivera-Batiz	A. R. Schotter
K. A. Mork	D. H. Perkins	R. Rob	S. L. Schreft
C. J. Morrison	J. M. Perloff	J. Roback	T. P. Schultz
D. T. Mortensen	P. Perron	D. J. Roberts	R. M. Schwab
R. B. Morton	T. Persson	S. Robinson	A. J. Schwartz
L. Moses	H. E. Peters	A. J. Robson	M. Schwartz
D. C. Mueller	P. Pfeiderer	D. Rodrik	G. W. Schwert
V. Munley	C. Phelan	C. A. Rogers	S. A. Scotchmer
K. Murphy	P. J. Pieper	W. P. Rogerson	L. Scott
K. J. Murphy	R. S. Pindyck	K. S. Rogoff	G. W. Scully
J. Mutti	D. Pines	L. Rojas-Suarez	J. J. Seater
J. H. Nachbar	C. A. Pissarides	R. Roll	R. A. Sedjo
B. Nalebuff	C. Pitchik	D. H. Romer	U. Segal
B. Naughton	C. R. Plott	P. M. Romer	G. Sellon
C. R. Nelson	I. Png	T. Romer	G. Shaffer
R. R. Nelson	S. W. Polachek	N. L. Rose	A. Shaked
P. Newbold	A. M. Polinsky	S. Rosefield	C. Shannon
J. P. Newhouse	W. Poole	S. Rosen	C. Shapiro
P. K. Newman	R. Porter	H. Rosenthal	M. D. Shapiro
D. Nichols	P. R. Portney	R. W. Rosenthal	S. Shavell

K. L. Shaw	R. M. Starr	M. A. Toman	L. H. White
S. M. Sheffrin	D. Starrett	R. H. Topel	L. J. White
T. Shen	R. Startz	B. Trehan	M. J. White
S. Shenker	J. C. Stein	R. Tresch	C. H. Whiteman
A. Shepard	R. Steinberg	R. L. Trosper	J. Whittaker
W. G. Shepherd	N. H. Stern	G. Tullock	R. Wigle
R. Sherman	S. Stern	S. Turnbull	D. W. Wilcox
R. J. Shiller	H. Stewart	A. Tversky	J. A. Wilcox
T. Shorrock	J. H. Stock	C. Udry	D. Wildasin
G. D. Short	A. C. Stockman	J. Underwood	L. L. Wilde
J. B. Shoven	T. M. Stoker	M. Ureta	S. R. Williams
M. Shroder	N. Stokay	D. Usher	S. D. Williamson
O. Shy	M. R. Stone	J. B. Van Huyck	R. J. Willis
D. S. Sibley	J. Strauss	R. A. Vanorder	J. D. Wilson
D. R. Siegel	M. E. Streit	B. van Velthoven	R. B. Wilson
J. Siegel	P. A. Streufert	H. R. Varian	S. Winer
H. A. Simon	M. H. Strober	W. K. Viscusi	R. A. Winter
L. Simon	C. Stuart	R. W. Vishny	R. S. Wintrobe
P. Sinclair	R. Sturm	X. Vives	M. Wiseman
K. J. Singleton	D. G. Sullivan	G. M. von Furstenberg	G. Woglom
J. Skinner	L. H. Summers	P. A. Wachtel	F. A. Wolak
S. Sklivas	S. Sunder	A. Wagstaff	B. L. Wolfe
M. E. Slade	K. Suzumura	M. Waldman	E. N. Wolff
J. Slemrod	L. E. Svensson	D. G. Waldo	A. Wolinsky
C. W. Smith	J. Sweet	J. Walker	K. I. Wolpin
R. S. Smith	D. Swenson	N. Wallace	S. A. Woodbury
R. T. Smith	J. M. Swinkels	J. J. Wallis	M. Wooders
V. K. Smith	S. Symansky	C. E. Walsh	M. Woodford
C. M. Snipp	G. Tabellini	G. Wang	R. Woods
J. M. Snyder	V. Tanzi	R. H. Webb	J. D. Worrall
J. Sobel	J. A. Tatom	S. B. Webb	G. Wright
G. R. Solon	P. Taubman	P. Weil	R. Wright
R. M. Solow	Y. Taumann	D. Weimer	L. B. Yeager
B. Sopher	G. Tavlas	B. R. Weingast	J. L. Yellen
F. Sowell	D. G. Taylor	E. R. Weintraub	J. M. Yinger
C. Spatt	L. D. Taylor	B. Weisbrod	A. Young
S. Spear	R. H. Thaler	A. Weiss	L. Young
B. J. Spencer	S. E. Thiel	W. Weissert	E. E. Zajac
M. W. Spicer	D. Thornton	M. L. Weitzman	G. A. Zarkin
K. E. Spier	M. Thursby	F. R. Welch	V. Zarnowitz
P. T. Spiller	T. N. Tideman	K. West	J. S. Zax
D. F. Spulber	T. Tietenberg	S. Weyers	R. Zeckhauser
D. O. Stahl	J. Tirole	W. C. Wheaton	S. P. Zeldes
R. W. Staiger	W. Todd	M. D. Whinston	P. Zemsky
O. Stark	R. D. Tollison	J. K. Whitaker	

ORLEY ASHENFELTER, *Editor*

Report of the Editor

Journal of Economic Literature

The mission of the *Journal* remains that of helping members of the Association maintain a broad understanding of developments in many branches of economics, and of assisting their work by providing them with a wide range of bibliographic material. These aims are served by articles, book reviews, annotations of new books, listings of the contents of current periodicals, selected abstracts, and listings of recent doctoral dissertations.

The *Journal's* work is divided between two offices. The Pittsburgh office under the direction of Drucilla Ekwurzel is responsible for the bibliographic material including the contents of current periodicals, certain abstracts, and new book annotations. Mary Kay Akerman is an Assistant Editor and Asatoshi Maeshiro is an Editorial Consultant to our Pittsburgh branch. They have been helped further by the work of Patricia Andrews, Elise Braden, Elizabeth Braunstein, and Elizabeth Thornton. I thank them all for their considerable efforts during the past year.

The Pittsburgh office is responsible also for the *Index of Economic Articles* and the *Economic Literature Index*. The latter is

available now not only on line through Dialog Information Services, but also in the form of compact discs through an arrangement with SilverPlatter Information. The EconLit database on disc provides over two decades of journal citations, abstracts of journal articles, book annotations, and dissertation titles. More information on computer access to the *JEL* bibliographic data is contained in each issue of the *Journal*.

Next year we shall be introducing a new Classification System for articles and books. We embarked on a revision of the old system two years ago and, having drawn upon the advice and suggestions of many people, we have been able to devise what we think is a scheme more appropriate to current economics. When we introduce this new Classification System, we shall merge our Subject Index of Articles in Current Periodicals with Selected Abstracts and thereby effect some economies. Though we try to contain the growth in the *Journal's* budget, the relentless expansion in economics publications makes this difficult. Table 1 provides information on the size of the *Journal's* various departments over the past decade.

TABLE 1—*JEL* PAGES BY DEPARTMENT, 1980–90

Year	Articles and Communications	Book Reviews	New Book Annotations	Current Periodicals ^a	General Index	Total
1980	366	294	276	1072	26	2034
1981	342	286	270	1059	23	1980
1982	331	251	300	1069	23	1974
1983	305	239	281	1086	38	1949
1984	354	225	314	1193	37	2123
1985	364	237	299	1306	38	2244
1986	326	250	308	1343	41	2268
1987	345	251	315	1352	40	2303
1988	419	241	318	1240	40	2258
1989	334	251	328	1254	41	2208
1990	323	234	366	1339	44	2306

^aIn 1987, the *Journal of Economic Literature* took over from the *American Economic Review* the responsibility of publishing the list of Doctoral Dissertations in Economics. This item is added to "Current Periodicals" which also includes the Contents of Current Periodicals, the Subject Index of Articles in Current Periodicals, and Selected Abstracts. The figures for 1990 include estimates regarding the December issue.

The Articles and Book Reviews departments of the *Journal* are handled by the Stanford office. During 1990, we published nine major articles, one review article, an exchange concerning the purposes of national income accounts, and 159 book reviews. Alex Field oversees the Book Review department and Moses Abramovitz assists me with the articles. Our work is supported by Anita Makler, Britt Ellis, and Toni Haskell. Ann Vollmer retired during the year and I wish to express my gratitude to her for her devoted service to the Association.

The expository and review articles appearing in the *Journal* are commissioned by the editor. However, I welcome proposals for such articles especially if such proposals are accompanied with a clear statement of the article's goals and an outline of its intended contents. All of our published articles go through a careful refereeing process.

I am most grateful to our referees for their very real assistance. I list at the end of this report the names of those referees who helped us during 1990 with manuscripts and proposed manuscripts. Many of these people wrote magnificent reports and made significant contributions to the quality of our articles.

The *Journal* has also benefited from the advice of its Board of Editors. Mark Rosenzweig's term on the Board will be completed by the end of 1990. I am very grateful to him for his careful work on behalf of the *Journal*. The terms of Richard Marston, Thomas Mayer, Robert Pollak, and Harvey Rosen are also coming to an end, but fortunately they have agreed to serve another term. Their service has been outstanding. I shall be proposing to the AEA Executive Committee that Sherman Robinson join the Board next year.

A. B. Abel	N. DeMarchi	N. M. Kiefer	Y. Qian
R. Arneson	D. Donaldson	C. P. Kindleberger	P. K. Robins
T. J. Bartik	M. Dotsey	R. C. Kormendi	L. B. Russell
B. W. Bateman	B. Eichengreen	J. R. Lave	W. F. Samuelson
C. Bell	R. Eisner	J. S. Leonard	I. V. Sawhill
A. Ben-Ner	R. J. Epstein	M. Machina	A. J. Schwartz
J.-P. Benassy	T. J. Espenshade	T. E. MaCurdy	M. D. Shapiro
T. C. Bergstrom	A. M. Feldman	W. G. Manning	T. M. Smeeding
B. Bernanke	M. A. Flavin	J. W. Marlin	J. P. Smith
B. D. Bernheim	M. Friedman	R. C. O. Matthews	V. K. Smith
D. S. Bizer	K. A. Froot	R. P. McAfee	R. W. Staiger
M. Blaug	R. E. Gallman	J. McMillan	R. M. Stern
D. Bloom	E. M. Gramlich	G. M. Meier	A. C. Stockman
J. H. Boyd	M. Gritz	A. H. Meltzer	D. A. Sumner
J. A. Brander	J. G. Gurley	P. Milgrom	M. D. Topper
G. F. Break	D. W. Hands	H. Miyazaki	S. J. Turnovsky
T. Bresnahan	E. A. Hanushek	R. A. Moffitt	D. W. Wilcox
J. P. Burkett	D. Hausman	T. A. Mroz	R. J. Willis
A. C. Cameron	W. P. Heller	R. J. Murnane	E. G. Winslow
R. S. Chirinko	C. Hildreth	J. M. Nason	F. Wolak
W. R. Cline	C. E. Hinshaw	R. G. Noll	A. Wolinsky
A. W. Coats	R. J. Hodrick	J. M. O'Brien	G. Wright
D. C. Colander	P. W. Howitt	J. D. Owen	
S. H. Danziger	F. M. Howland	J. K. Peskin	
P. Dasgupta	T. Josling	R. S. Pindyck	

JOHN PENCANEL, *Editor*

Report of the Editor

Journal of Economic Perspectives

The end of 1990 marks three-and-one-half years of publication of the *Journal of Economic Perspectives*. During that time, the *Journal* has been evolving from complete newcomer to familiar face. As far as the editors can judge, the first 14 issues of the *Journal* have developed a reputation for interest and clarity, and a loyal ongoing readership.

The *Journal* has continued the pattern set in years past of addressing a broad range of topics in several different formats. The issues of 1990 included 6 symposia: new institutions for developing country debt; welfare and workfare; bubbles; the state and economic development; collaboration, innovation, and antitrust; and the economic status of African-Americans. They also included 4 articles about well-known economists (Houthakker, Kreps, Pechman, and Goldberger), and 12 articles on various other topics, ranging from bimetallism to socialist economics, from micro-macro simulation to the econometrics of kinked budget constraints, from optimal taxation to state lotteries, and others. In addition, the *Journal* also continued publication of several regular features—"Puzzles," "Anomalies," "Recommendations for Further Reading," along with Notes and Correspondence.

Two new features began in 1990, as well. "Retrospectives" will be aimed at topics in the history of economic thought, while "Policy Watch" will deal with policy topics of immediate relevance. In 1991, these new features will partly replace the "Puzzles" and "Anomalies" columns.

In total, the 1990 issues of the *Journal* included 896 pages, consisting of 36 articles, 5 introductions or very brief articles, 16 of the features articles, plus Notes and Correspondence. In the two previous years, the *Journal* had published a total of 832 pages, so the number of journal pages increased by about 8 percent in 1990. This was mainly due to an increase in the short feature arti-

cles from 11 in 1989 to 16 in 1990. As we publish fewer columns of puzzles and anomalies in 1991, the 1991 page total should decline from the 1990 level.

The decentralized editorial procedure of the *Journal* has continued to work well, as the list of articles illustrates. The Associate Editors have continued to generate lists of ideas and potential authors, enabling the *Journal* to draw on a large group of economists with diverse interests and perspectives. A process of natural turnover seems to have developed among the Associate Editors, allowing several new members to bring their ideas to the *Journal* each year.

Many readers have let us know that they would like to see some articles about the recent changes in the Soviet Union and Eastern Europe. One of the new Associate Editors for 1990 (Peter Murrell of the University of Maryland) will take on the responsibility of soliciting articles about these topics. The *Journal* plans to publish a lengthy symposium on these events in its Fall 1991 issue, and then several follow-up articles in future years.

As in years past, the *Journal* has had no difficulty in finding authors who were willing and eager to write. The broad readership of the *Journal* seems especially attractive to potential authors and also makes them willing to work hard on revising and reshaping their articles in order to make them accessible to that broad audience. As a result, the issues of 1991 are almost fully booked with articles of high quality, and advance planning for 1992 and 1993 is well underway.

The incumbent Associate Editors who will continue into 1991 are: Henry Aaron, The Brookings Institution; Pranab Bardhan, University of California-Berkeley; James Heckman, Yale University; Dwight Jaffee, Princeton University; N. Gregory Mankiw, Harvard University; Barry Nalebuff, Yale University; John Roemer, University of Cal-

ifornia-Davis; Sherwin Rosen, University of Chicago; Bernard Saffran, Swarthmore College; Steven Salop, Georgetown University; Hal Varian, University of Michigan; Gavin Wright, Stanford University; Janet Yellen, University of California-Berkeley.

Members of the *Journal's* Advisory Board have continued to provide helpful advice, including feedback on past articles and suggestions for future ones. The *Journal* has continued its tradition of holding a breakfast each year at the annual meetings, so that members of the Advisory Board, the Associate Editors, and the Editors can meet to brainstorm and exchange views about the *Journal*.

The staff of the *Journal* remained the same in 1990. Debbie Sachs held the position of Editorial Associate. The day-to-day job of running a journal requires a considerable array of personal, management, and administrative skills. Debbie's flexibility, good humor, and attention to detail have

been essential to the smooth functioning of the *Journal*.

As in years past, the Editors feel that they cannot overstate the role that Managing Editor Timothy Taylor has played in the operation of the journal. He has managed the day-to-day operations of the *Journal* smoothly and ensured that the fundamental objectives of the *Journal* are satisfied. He has performed the difficult task of persuading authors to amend and rewrite their articles with delicacy and skill, and has shown that it is possible to edit papers and increase their clarity and accessibility, while still retaining the distinctive voice of each author.

Requested action: 1) Approval of new members to join the Editorial Board in 1991: Peter Murrell, University of Maryland; Gene Grossman, Princeton University.

JOSEPH STIGLITZ, *Editor*
CARL SHAPIRO, *Co-editor*

Report of the Director

Job Openings for Economists

The total number of new jobs listed decreased from 2,014 last year to 1,875 this year. Academic jobs increased from 1,411 to 1,447, but nonacademic listings declined to 428 from 603. Table 1 shows total listings (employers), total jobs, new listings, and new jobs by type for each issue of *JOE* in 1990.

Table 2 shows the number of employers by category (four-year colleges, universities with graduate programs, federal government, etc.) for each of the seven issues. Academic institutions continue to be the major advertisers, about 77 percent of the total number of employers listing vacancies.

TABLE 1—JOB LISTINGS FOR 1990

Issue	Total Listings	Total Jobs	New Listings	New Jobs
Academic				
February	76	155	66	132
April	48	94	44	84
June	28	61	26	53
August	63	146	58	135
October	164	447	151	423
November	156	310	156	310
December	153	367	92	203
Subtotal	688	1,570	593	1,447
Nonacademic				
February	13	55	10	37
April	18	58	13	36
June	20	72	15	51
August	16	45	13	29
October	49	160	43	132
November	26	56	26	56
December	57	205	25	87
Subtotal	199	651	145	428
Total	887	2,221	738	1,875

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1990

Issue	Four-Year Colleges	Universities with Graduate Programs	Federal Government	State/Local Government	Banking or Finance	Business or Industry	Consulting or Research	Other	Total
February	28	48	2	—	3	1	6	1	89
April	18	30	3	1	3	2	8	1	66
June	11	17	5	1	2	1	11	—	48
August	21	42	1	—	8	—	6	1	79
October	49	115	14	5	6	2	18	4	213
November	63	93	3	1	6	2	11	3	182
December	49	104	18	1	9	2	21	6	210
Totals	239	449	46	9	37	10	81	16	887

TABLE 3—FIELDS OF SPECIALIZATION CITED: 1990

Fields	February	April	June	August	October	November	December	Totals
General Economic Theory (000)	55	42	26	53	200	163	200	739
Growth and Development (100)	24	19	5	5	39	23	34	149
Econometrics and Statistics (200)	18	21	9	14	78	51	73	264
Monetary and Fiscal (300)	30	14	15	25	95	67	95	341
International Economics (400)	30	21	17	23	67	62	79	299
Business Administration, Finance, Marketing and Accounting (500)	33	19	8	13	45	34	52	204
Industrial Organization (600)	25	18	14	14	82	37	86	276
Agriculture and Natural Resources (700)	11	10	9	16	33	16	38	133
Labor (800)	17	5	6	13	46	37	55	179
Welfare and Urban (900)	15	6	9	18	76	34	55	213
Related Disciplines (A00)	3	1	—	2	10	4	7	27
Administrative Positions (B00)	11	4	6	5	13	9	11	59
Totals	272	180	124	201	784	537	785	2,883

Note: Fields of specialization codes are from the *Journal of Economic Literature*.

The distribution of employers by type has been roughly the same for the past several years.

Table 3 shows the number of listings by field of specialization. Once again, General Economic Theory (which includes both micro and macro) led in popularity, and Monetary and Fiscal (300) was second. International Economics (400) beat out Econometrics and Statistics (300) for third—a change in the usual pattern that has prevailed over many years. The new classi-

fication scheme of the *JEL* will be utilized next year. The new categories, especially the breaking up of General Economic Theory into more detailed specialties, will reveal more about which fields are “hot.”

Violet Sikas continues to be the “one-woman band” for *JOE*. Without her, the Director might have to do some work. I am grateful for all that she does.

C. ELTON HINSHAW, *Director*

Report of the Committee on Economic Education

The third edition of the *Test of Understanding College Economics* (TUCE III) is now available from the Joint Council on Economic Education, 432 Park Avenue South, New York, New York 10016. The TUCE III consists of a 33-item, four-option multiple-choice test in macroeconomics and a 33-item, four-option multiple-choice test in microeconomics. The last three questions on each test deal with international economics concepts. This enables instructors who do not cover international economics in the macro course (or in the micro course) to use the first 30 questions as a "pure" macro (or micro) test.

The overall content and cognitive specifications for TUCE III are similar to those of previous editions, but only three of the questions on the macro test and only four of the questions on the micro test have exactly the same wording in TUCE III as in TUCE II. A set of macro norming data collected from 92 sections at 46 institutions and a set of micro norming data collected from 96 sections at 49 institutions is reported and interpreted by Phillip Saunders of Indiana University in the *Interpretive Manual* that accompanies TUCE III. Saunders is in the process of compiling an extensive set of student and instructor information provided by the norming institutions. When completed, this data set should be of considerable interest to researchers who want to analyze and interpret student performance on TUCE III in a variety of different institutions. Like TUCE itself, the *Interpretive Manual* and the forthcoming data set will be available from the Joint Council on Economic Education.

The Association of American Colleges (AAC) has completed its project on study-in-depth. The AEA Committee on Economic Education participated in the project along with eleven other arts and science disciplines. A task force of six economists (all but one of whom are current members or former chairs of the Committee) has completed an extensive report on the eco-

nomics major. The report documents the number of students who have majored in economics over the last two decades, attempts to describe a consensus view of the purpose of the major, describes the typical curriculum, examines several pedagogical issues (class size, instructional methods, writing, perspective on the methodology of economics, and the effect of teaching methods on students of different ages, race, and gender), concludes that we know little about what our students learn in the major, and makes a series of recommendations to improve the experience of students who major in economics. The complete report is scheduled to appear in the Summer 1991 issue of the *Journal of Economic Education* along with comments by two discussants. A shorter version appears along with essays on the major in eleven other disciplines in a volume published by the AAC. A summary of the paper along with the recommendations is part of the proceedings of the December 1990 AEA meetings.

There has been considerable research on high school economics in recent years. A grant from the Pew Trust to the Joint Council on Economic Education in the late 1980's nurtured this growth. The Economics Education Session sponsored by the Committee on Economic Education at the December 1989 AEA meetings focused on the new research results which emanated from the Pew Project. To build on this momentum, the Committee recommended preparation of a survey of the research literature on high school economics similar to the survey of research on teaching college economics that was published in the *Journal of Economic Literature* in 1979. Work on the survey article is underway.

The Committee devoted a special meeting in 1989 to establishing an agenda for its work in the 1990's. A subcommittee produced an agenda statement, which is part of the *Proceedings* of the December 1990 AEA meetings. The Committee plans to focus on three areas: 1) Expanding and improving

our ability to assess learning (particularly above the principles level, and especially with instruments other than multiple-choice tests); 2) Learning more about the economics major; and 3) Developing data sets and encouraging researchers to reexamine

many of the important empirical issues in economics education with improved data sets and modern empirical methods.

JOHN J. SIEGFRIED, *Chair*

Report of the Committee on the Status of Minority Groups in the Economics Profession

The primary mission of the Committee on the Status of Minority Groups in the Economics Profession (CSMGEP) is to increase the representation of minority groups in economics. The targeted groups are those that are underrepresented within the field, particularly African-Americans, Hispanics, and Native Americans. The Committee has had three organized programs to facilitate the achievement of this objective—the AEA Summer Minority Program, the AEA/Rockefeller graduate fellowship program, and the AEA/FRS dissertation fellowship program. Together they have provided supplementary preparation for graduate study and the financial support that would allow students to complete their Ph.D.s in economics. During 1990, only the summer program and the dissertation fellowships were in operation and significant changes were made in both of them.

AEA Summer Minority Program. The summer program is designed to increase the probability that promising minority undergraduate students enter and complete Ph.D. programs in economics. It has been in operation since 1974, providing minority undergraduate students (rising seniors) with an eight-week program of theoretical and applied economics. This year marked the fifth and final year for Temple University to serve as host for the junior summer program.

In 1989, the CSMGEP solicited applications from universities to replace Temple as the host institution and, by the beginning of 1990, four completed applications were received. All four institutions indicated a willingness to commit substantial resources, over and above the support to be provided by the AEA. Stanford University was selected as the new host for the summer program, and they have made information on the program available through advertisements and direct mailings. The principal features of the Stanford program include a highly analytical curriculum, course credit and transcripts for students completing the

courses, and weekly policy seminars. The program will be directed by Thomas MaCurdy of the Economics Department, but includes participation by both the Food Research Institute and the Graduate School of Business at Stanford.

AEA / FRS Fellowship. The Federal Reserve System has sponsored a program in conjunction with the AEA which provides financial support to the third, fourth, and possibly fifth-year minority graduate students who are writing dissertations in areas of interest to the Federal Reserve. This support has been renewed for a five-year period, beginning with the 1991–92 academic year. The stipend has been raised to \$900 per month for the academic year, and it includes a requirement that students spend the summer after their first fellowship year as an intern at the Board or at one of the Federal Reserve Banks. Up to ten awards will be made each year.

During the 1990–91 academic year, support is being provided to Grover McArthur at The New School for Social Research, Victor Mendes at Syracuse University, and Gregory Price at the University of Wisconsin-Milwaukee. Among the group of recipients during the 1989–90 academic year, Philips Jefferson completed his Ph.D. at the University of Virginia.

Other Activities. The Committee has the obligation of seeking funding to replace the AEA/Rockefeller fellowship program, which provided support for the first and second years of graduate study. Sources of funding were identified in 1990 and proposals will be developed and submitted in 1991. The development of a program that would increase interest in economics among college sophomores is being considered, possibly in conjunction with other organizations, such as the National Economic Association.

A survey of past participants in the summer program was conducted in 1990 and results are being tabulated. In the future, more frequent surveys of program partici-

pants will be carried out, in conjunction with the program's host institution.

On July 1, Ronald Oaxaca completed his term as Chair of CSMGEP. William A.

Darity, Jr. ended his second term on the Committee at the end of the year.

MARGARET C. SIMMS, *Chair*

Report of the Committee on the Status of Women in the Economics Profession

The Committee on the Status of Women in the Economics Profession (CSWEP) has been charged by the American Economic Association with monitoring the position of women in the profession and undertaking activities to improve it. This Report examines the advancement of women economists in academia, compares this progress with what might be expected, and describes the Committee's activities during the past year.

Are Women Economists as Likely as Men to be Hired and Promoted?

The proportion of assistant professors who are women has been rising as one would expect based on the growing proportion of Ph.D.s in economics awarded to women, but their progress into the ranks of associate and full professor appears to be lagging somewhat. This conclusion is based primarily on data about graduate economics departments (defined as those that award Ph.D.s) that responded to the AEA's Universal Academic Questionnaire (UAQ) between 1974 and 1989.¹

Two parallel analyses of full-time faculty were conducted for this Report: One used data from 150 graduate economics departments that responded to the questionnaire in any year, while the other examined the 43 departments that responded in almost every year.² Because of the similarity in results, only the analysis for the larger sample of departments is reported here.

The proportion of assistant professors and of associate professors who were women approximately tripled between 1974 and 1989—rising from 8 to 20 percent and from

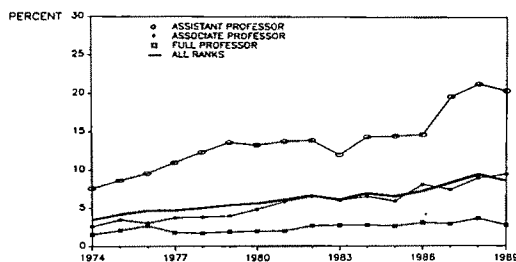


FIGURE 1. PROPORTION OF FACULTY IN
GRADUATE DEPARTMENTS WHO ARE WOMEN
By Rank, 1974-89

3 to 9 percent, respectively (see Figure 1).³ In contrast, the proportion of full professors who were women grew from about 2 percent in the late 1970's to about 3 percent in the late 1980's. Because about one-half of all tenure-track and tenured economics faculty at departments in the sample were full professors, the proportion of all faculty who were women was about 9 percent in 1989. This represented a doubling of the proportion from 1970's.

Differences in the employment of women faculty by different types of institutions have not been dramatic. Women assistant professors of economics were somewhat more likely to be employed by public institutions in the 1980's than men, while there was little difference between public and private institutions for associate or full professors (see Figure 2). When departments are ranked by the scholarly quality of their faculty, it appears that women assistant professors might have been employed by lower-quality schools relatively more often than men during the 1980's; whereas, there was

¹The Committee thanks Eric Guille, Jodi Korb, Charles Scott, Arantza Ugidos, Jackie Vander Brug, and Bruce Vavrichek for their contributions to this Report.

²In both cases, departments of agricultural economics were excluded. The 43 departments are those that did not respond in at most two consecutive years.

³In this analysis, the proportion who are women is always of the comparable group of faculty—in this case, it is the percent of faculty with the same rank who are women. (The source for all figures herein is the UAQ.)

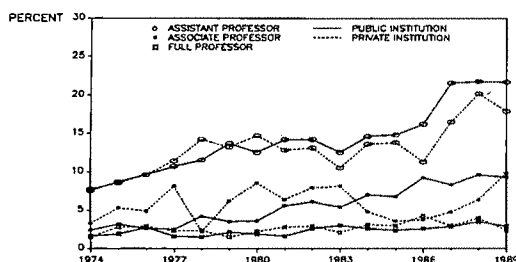


FIGURE 2. PROPORTION OF FACULTY IN GRADUATE DEPARTMENTS WHO ARE WOMEN BY RANK AND TYPE OF INSTITUTION, 1974-89

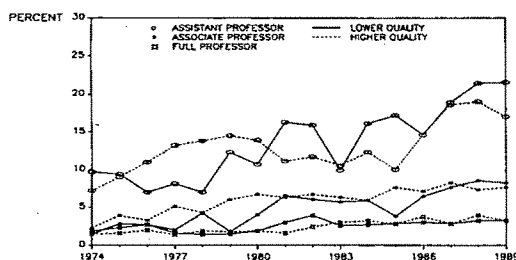


FIGURE 3. PROPORTION OF FACULTY IN GRADUATE DEPARTMENTS WHO ARE WOMEN BY RANK AND QUALITY OF DEPARTMENT, 1974-89

little difference for associate or full professors, as shown in Figure 3.⁴

How does the progress of women through the academic ranks compare with what would be expected if hiring and promotion decisions were unaffected by gender? Figure 4 shows that the proportion of newly hired assistant professors who were women has risen as expected—it resembled the proportion of new recipients of Ph.D.s in economics who were women, at least until late in the 1980's.⁵

⁴The scholarly quality of economics departments was based on data for 93 graduate economics departments reported in National Research Council, *An Assessment of Research-Doctorate Programs in the United States: Social and Behavioral Sciences* (Washington: National Academy Press, 1982). The data examined here included 88 of those departments; the remaining 62 departments (of the 150 on which this report is based) that were not included in the NRC's study were excluded from the calculations for Figure 3.

⁵Data on newly awarded Ph.D.s in economics are reported in National Science Foundation, *Science and*

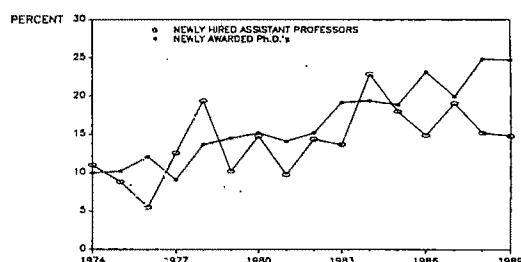


FIGURE 4. PROPORTION OF NEWLY HIRED ASSISTANT PROFESSORS IN GRADUATE DEPARTMENTS, AND PROPORTION OF NEWLY AWARDED PH.D.S IN ECONOMICS WHO ARE WOMEN, 1974-89

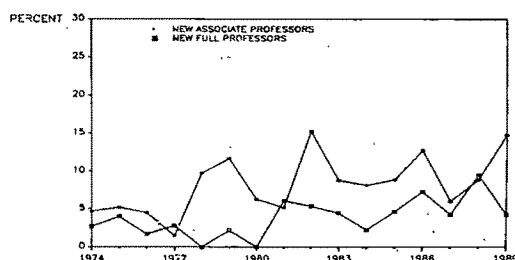


FIGURE 5. PROPORTION OF NEW ASSOCIATE AND FULL PROFESSORS IN GRADUATE DEPARTMENTS WHO ARE WOMEN, 1974-89

The proportion of newly hired or newly promoted associate (and full) professors who were women has also been rising, as shown in Figure 5, but assessing this growth is more difficult because, a priori, it is not clear what the standard of comparison should be. To deal with this problem, actual data were compared with the results of a simple model that simulates the promotion process in academia. The assumptions underlying the model were determined by data about the flow of faculty into each rank (i.e., newly hired individuals and those promoted from the lower rank), and about the flow

Engineering Doctorates: 1960-1989. They are based on all graduate departments on the United States. In Figure 4, data on newly awarded Ph.D.s are for U.S. citizens and noncitizens who are permanent residents. Temporary residents are excluded.

out of each rank (i.e., faculty who left the department and those who were promoted into the next rank).⁶ In particular, the model assumes that, on average, full professors remain in a department for 20 years; that promotion to full professor is decided, on average, 7 years after promotion to associate professor; that promotion to associate professor is decided, on average, 5 years after being hired as an assistant professor; and that women are hired as assistant professors in the same proportion as they receive newly awarded Ph.D.s in economics.⁷ In addition, the model assumes that the distribution of talent and motivation is the same for female economists as for male economists and, hence, that the probability of an individual being promoted is not related to gender.

Figure 6 compares the implications of the model with what actually happened. The proportion of assistant professors who were women increased as the model projects over the 1974–89 period, with the actual proportion exceeding the projected one in the first half of the period and falling below for much of the second half. In contrast, the actual proportions of associate professors and full professors who were women have been below the levels projected by the model in most years and consistently below since 1982.

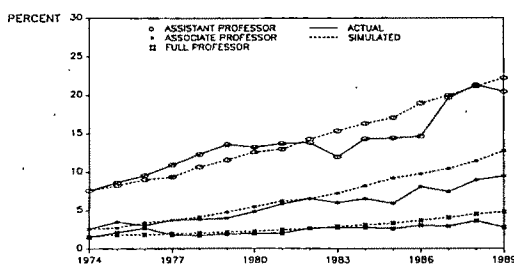


FIGURE 6. ACTUAL AND SIMULATED PROPORTION OF FACULTY IN GRADUATE DEPARTMENTS WHO ARE WOMEN, BY RANK, 1974–89

These results can be seen as both discouraging and encouraging. Unfortunately, they support the belief that women's progress to the ranks of associate and full professor has been slower than it should have been. More positively, though, they indicate that any shortfall (especially in the proportion of full professors who are women) is considerably less than might be assumed based on the much faster growth in the proportion of assistant professors who are women. In other words, that simple comparison neglects the fact that, on average, it takes more than a decade to advance from assistant to full professor.

What might we conclude about the future? On the one hand, the rising proportion of assistant professors of economics who are women indicates that there is a growing pool of women to be promoted to associate and full professor. On the other hand, if women assistant professors are disproportionately located at lower-quality departments, growth in the proportion of associate and full professors who are women may be limited at the better schools. Moreover, the actual increase in the proportion of full professors who are women over the next decade will be largely constrained by what has already happened for associate professors. Yet, the actual proportion of associate professors who are women has been noticeably below the simulated proportion since 1982. Thus, the gap between the actual and the simulated proportion of full professors who are women is likely to grow in coming years. While future research

⁶For each year, the percentage of faculty in each of these categories was averaged over the departments that responded in that year. Because these percentages were reasonably stable over the 1974–89 period, the 16-year averages were used in the simulation model.

⁷In 1975, for example, the model assumes that the proportion of faculty leaving the rank of full professor who were women depended on the proportion of doctoral recipients who were women in 1943 (i.e., 32 years earlier because these faculty, on average, spent 20 years as full professors and took 12 years to reach that rank after being hired as assistant professors). Because actual data on doctoral reciprocity were not available for years prior to 1960, the model assumes that the proportion awarded to women was 4 percent in each earlier year—about the same as in the early 1960's. This assumption is conservative (i.e., it tends to lower the projected proportion of full professors who are women) because doctoral reciprocity in the early years affects only the simulated outflow of women full professors.

may be able to shed some light on why these patterns persist, expanding women's participation in the economics profession may also require examining the interpersonal dynamics within graduate programs and departments.

The Committee's Recent Activities

CSWEP pursued several activities in 1990 designed to help women advance in the economics profession. As part of its ongoing efforts to expand the participation of women economists on the program of the AEA's annual meetings, the Board organized six sessions for 1990—three on gender-related topics and three on technology and productivity. In addition, each person asked by the President-elect to organize an invited-paper session was sent a list of experienced women economists who specialize in the same fields as the organizer. These reminders of possible participants have been associated with greater participation of women.⁸ As well, to facilitate networking at the annual meeting among economists who support CSWEP's goals, CSWEP sponsored a hospitality suite and a reception followed its business meeting.

Another major activity was publishing three issues of the CSWEP *Newsletter*, the contents of which are designed primarily to help young economists advance. Each issue contains information about sources of research funding and calls for papers, as well as articles on topics such as the annual job market and tips on writing publishable articles. In response to many requests for copies of articles from earlier issues, the Board also reprinted selected articles in a fourth

issue of the *Newsletter* that is available free to dues-paying members, or for \$8.

Updating and expanding the entries in the *Roster of Women Economists* absorbed a considerable amount of the Board's resources. The *Roster* contains information about women economists, including their employers, educational backgrounds, fields of specialization, and publications. It is used by employers searching for job candidates and by organizations seeking members of advisory committees and the like. The entire *Roster*, or selected portions of it, are available in computer-readable form or as mailing labels. In addition, the *Roster* appears in a printed volume every other year. The Board also continued its recent practice of informing advertisers in *Job Openings for Economists* and the CSWEP *Newsletter* about the *Roster* and how to use it.

Finally, the Board thanks Joan Haworth, the Committee's Membership Secretary, and her staff for their many contributions—maintaining the *Roster*, updating it using the AEA's membership directory, preparing special mailings, and creating customized listings from the *Roster*, to name just a few. The Board is also grateful to three members whose terms expire this year. Shulamit Kahn organized many sessions and coordinated information about the availability of child care for several annual meetings of the AEA. She has also agreed to continue participating in an ongoing CSWEP-sponsored research project comparing the career paths of female and male economists. Kathryn Morrison and Barbara Newell each co-edited an issue of the *Newsletter* and helped with projects to expand use of the *Roster*. The Board also thanks Jill Bury, who continues to contribute a great deal, including doing an outstanding job producing the *Newsletter*.

NANCY M. GORDON, *Chair*

⁸See Cecilia A. Conrad, "Women Economists at the AEA Annual Meeting," Barnard College, Columbia University, October 1990.

Report of the Committee on U.S.-China Exchanges in Economics

Modern economic education and research continued to develop in China in 1990 as evidenced by the following events.

1) American economists continued to teach in China, including the two graduate economics training centers at the People's University in Beijing (Harold Watts of Columbia University and Kajal Lahiri of the State University of New York at Albany in Spring 1990; Mark Machina of the University of California-San Diego and Belton Fleisher of Ohio State University in Fall 1990) and at Fudan University in Shanghai (Alastair MacBean of Lancaster University and Sumner LaCroix of the University of Hawaii in Spring 1990; Dudley Wallace of Duke University and Albert Schweinberger of Australian National University in Fall 1990), and Lingnan (University) College at Zhongshan University in Guangzhou (Anthony Koo of Michigan State University).

2) A delegation of deans and chairmen of colleges and departments of economics headed by the Dean of Economics at the People's University, Yu Xueben, visited the United States on October 27 to November 18 to study economic education in the United States. The seven major universities represented were Fudan, Jilin, Nankai, Peking, the People's, Wuhan, and Xiamen universities. Besides visiting Columbia, Harvard, Washington, D.C., and Stanford, the delegation spent ten days at the University of Michigan, hosted by Michael Oxenburgh of the Center for Oriental Studies and Robert Dernberger of the Department of Economics, and participated in intensive discussions on the administrative and academic aspects of American economic education. Other visitors from China included visiting fellows supported by The Ford Foundation. Dr. Justin Lin of the Rural Development Department, Development Research Center of the Chinese State

Council and Beijing University, served as Visiting Associate Professor at the University of California-Los Angeles, in the fall quarter 1990.

3) An International Conference on Quantitative Economics and Its Applications to Chinese Economic Development and Reform in the 1990's, sponsored by the Chinese Academy of Social Sciences and the United Nations Development Programme in China, took place in Beijing, June 24-28. Sixteen foreign economists participated, mostly American, but including French, Hungarian, and Taiwanese economists. The conference was to celebrate the tenth anniversary of the Econometrics Workshop organized by Lawrence R. Klein of the University of Pennsylvania and Xu Dixin of the Chinese Academy of Social Sciences in Beijing in the summer of 1980. The instructors of the 1980 workshop, including Albert Ando, Gregory Chow, Cheng Hsiao, Lawrence Klein, Lawrence Lau, and Vincent Su returned to Beijing to present papers and to visit the site of the 1980 workshop. A number of papers were presented by Chinese scholars, including those of the Chinese Academy of Social Sciences doing cooperative research with their American colleagues.

4) The Chinese State Education Commission continued to develop modern economic education by introducing the core courses of micro, macro, statistics, accounting, international trade, economic development, and money and banking, which were adopted in 1988. In the summer, a number of workshops were organized to train the university teachers to teach some of these core courses. Winston Chang of the State University of New York at Buffalo and Frank Hsiao of the University of Colorado were among the visiting professors teaching the workshops.

GREGORY C. CHOW, *Chair*

Report of the Representative to the National Bureau of Economic Research

The National Bureau of Economic Research studies a wide variety of economic issues. Much of this research is conducted by economists working individually, but a large part is organized into special projects. To disseminate the results of this research, during 1990 the NBER issued almost 400 working papers, published 10 books, 2 NBER journals, prepared 6 special issues of other journals, and circulated the monthly *Digest* and the quarterly *Reporter*. In addition, the NBER held numerous conferences and workshops, including the 6-week long Summer Institute. Finally, the NBER hosted three Olin Fellows and two Aging Fellows who spent the year conducting empirical research on a variety of empirical topics.

Programs. The NBER's ongoing programs generally meet twice during the academic year, and once, for a longer period, during the Summer Institute. Bureau programs (directors in parentheses) are Economic Fluctuations (Robert Hall), Financial Markets and Monetary Economics (Benjamin Friedman), International Studies (William Branson), Labor Studies (Richard Freeman), Taxation (David Bradford), Development of the American Economy (Robert Fogel and Claudia Goldin), Health Economics (Victor Fuchs and Michael Grossman), and Productivity (Zvi Griliches). In 1990, Nancy Rose became the director of a new Bureau program in Industrial Organization.

Projects. The NBER's projects generally bring a dozen or more researchers together to work on a common topic. The projects' findings are usually distributed initially as NBER working papers, and final versions are then published as Bureau books. During 1990, the following projects (organizers in parentheses) were underway: Trade and Immigration (George Borjas and Richard Freeman), Economic Growth (Robert Barro and Paul Romer), Economics of Higher Education (Charles Clotfelter and Michael Rothschild), Financial Crisis

(Glenn Hubbard), Tax Reforms in Asia and the U.S. (Anne Krueger and Takatoshi Ito), Macroeconomic History (N. Gregory Mankiw and Christina Romer), Politics and Economics in the 1980's (Alberto Alesina and Geoffrey Carliner), Exchange Rate Targets and Currency Bands (Paul Krugman and Marcus Miller), Trade Policy and the Uruguay Round (Robert Baldwin and David Richardson), Taxation in the U.S. and Canada (John Shoven and John Whalley), Macro Policies and Income Distribution in Latin America (Rudiger Dornbusch and Sebastian Edwards), Economics of Aging (David Wise), and American Economic Policy (Martin Feldstein).

Other Conferences. In addition to the conferences and workshops conducted as part of NBER projects and programs, the NBER also holds larger conferences open to individuals in government and academe. During 1990, these included a session of the Conference on Income and Wealth on the Measuring Service Sector Output, organized by Zvi Griliches; a Universities Research Conference on Asset Pricing and Financial Markets, organized by John Campbell; and a second URC on Exchange Rate Regimes, organized by Alberto Giovannini. Procedures and deadlines for submitting papers to these conferences are posted in economics departments and announced in the *NBER Reporter*.

Summer Institute. During the summer of 1990, 642 economists from 180 different universities and research organizations attended the Summer Institute, including 255 participants who came for the first time. Participants stayed for periods ranging from 4 days to 2 months. There were 273 papers presented in 36 separate sessions. A catalogue of these papers, including authors and their addresses, is available from the NBER on request.

Working Papers. NBER working papers are circulated to economics department libraries around the world, and are available

from the NBER for \$3 per title in the United States and \$4 elsewhere. Subscriptions to the entire series are available for \$500 per year inside the United States and slightly more elsewhere.

Books. The 10 books published by the NBER during 1990 included monographs by Robert Gordon on measuring durable goods' prices and by Robert Margo on race and schooling in the South, 1880–1950. Assaf Razin and Joel Slemrod edited a volume on international taxation. William Branson, Jacob Frenkel, and Morris Goldstein edited a volume on international policy coordination and exchange rate fluctuations.

Journals. The NBER publishes two journals, which appear once each year. Olivier Blanchard and Stanley Fischer edited the *Macroeconomic Annual* in 1990, and Lawrence Summers edited *Tax Policy and the Economy*. Subscriptions for these two journals can be ordered from MIT Press.

In addition, the NBER sponsored special issues of several journals. Tony Atkinson and David Bradford edited papers on the future of the welfare state for the *Journal of Public Economics*. Sanford Grossman organized the *Review of Financial Studies* symposium on stock market volatility. Robert Gordon and Georges de Menil edited papers on international macroeconomics for

the June issue of the *European Economic Review*. Edmar Bacha and Sebastian Edwards edited papers on Latin America for publication in the *Journal of Development Economics*. John Campbell and Angelo Melino organized the issue of the *Journal of Econometrics* on financial times-series.

Periodicals. The monthly NBER *Digest* provides brief summaries of working papers of general interest. The quarterly *Reporter* contains longer summaries of recent research, abstracts of all working papers, announcements of the publication of NBER books, and a calendar of forthcoming conferences. Both these periodicals are available on request.

Olin Fellows. The three Olin Fellows for the 1990–91 academic year are Lawrence Ball, Robert Gibbons, and Dani Rodrik. These fellowships are open to young empirical economists at any academic institution. The two Aging Fellows are Ed Norton and Leslie Papke.

During 1990, Martin Feldstein continued as President of the NBER and Geoffrey Carliner continued as Executive Director. Further information on NBER activities is available in the *Reporter*, or from Geoffrey Carliner, NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138.

DAVID KENDRICK, *Representative*

Report of the Representative to the American Association for the Advancement of Science

The American Association for the Advancement of Science is a federation of scientific organizations, as well as an association of over 130,000 individual members. Its objectives are "to further the work of scientists, to facilitate cooperation among them, to foster scientific freedom and responsibility, to improve the effectiveness of science in the promotion of human welfare, and to increase public understanding and appreciation of the importance and promise of the methods of science in human progress."

The discipline of Economics is grouped with Political Science, Sociology, Geography, and some related social sciences in Section K, one of the twenty-one divisions within the AAAS. Most economists list Section K as their primary affiliation, and some hold dual or multiple membership in other sections. Overall, the four social science sections comprise just under 10 percent of the AAAS-affiliated membership.

The AAAS offers economists a unique opportunity to interact with the scientific community at large. This includes the following channels: the AAAS annual meeting, joint sessions with other organizations at their professional meetings, speciality conferences, *Science*, and AAAS-sponsored books. In addition, the organization is active through the work of its offices and standing committees on Science and Technology Education, Public Sector Programs, Opportunities in Science, International Science, Scientific Freedom and Responsibility, and Science, Arms Control, and National Security.

A joint AEA/AAAS session was held at the 1990 Allied Social Science meetings on "R&D Spillovers and Externalities." The session was chaired by Richard Levin and included papers by Jess Benhabib and Boyan Jovanovic, Jeffrey Bernstein and M. Ishaq Nadiri, Edwin Mansfield, and Adam Jaffee and Rebecca Henderson.

Three of the symposia and five of the technical sessions at the 1990 AAAS meet-

ing in New Orleans received the endorsement of the AEA/AAAS Liaison Committee. In addition, AEA members organized or contributed to other sessions. Economists participating in the annual AAAS meeting included John Antle, Gilbert Bassett, Robert Evenson, Donald Hanson, Richard Kosobud, Lester Lave, William Nordhaus, Mancur Olson, James Opaluch, Paul Portney, Clifford Russell, Roger Sedjo, James Smith, Andrew Solow, and Gary Yohe. Session topics covered a diversity of topics such as Foreign Assistance and Environmental Problems in LDCs, Economic Perspectives on Biotechnological Development, Managing Outer Continental Oil and Gas Resources, and Agricultural Chemicals and Water Quality.

Through the efforts of the AEA/AAAS Liaison Committee and the AEA membership, economists have been more active than other social scientists at the AAAS annual meeting. Our continued success at getting sessions devoted to economics on the program is dependent on the submission of high quality proposals and strong attendance. AEA members are encouraged to take part at future AAAS meetings, the next one to be held in Chicago in February 1992.

Continuing the trend of the past 2 years, several economists were commissioned to write papers for *Science*. Authors of commissioned or contributed papers included: Victor Fuchs, John Kennan, Robert Litan, Ronald Mincy, Alvin Roth, Isabell Sawhill, Peter Suchman, W. Kip Viscusi, Robert Wilson, Douglas Wolf, and Richard Zeckhauser. Robert Solow and Elizabeth Bailey continue to serve on the Editorial Board of *Science* and Robert Dorfman on the publication's Board of Reviewing Editors.

Several AEA members are officers of Section K. They include: Mancur Olson (Retiring Chair), Marc Nerlove (Chair-Elect), Robert Fogel and Amitai Etzioni (Section Committee Members), Richard Cyert, Lester Lave, and T. Paul Schultz

(Electorate Nominating Committee Members), and William Alonso (Council Delegate).

More than 30 economists are fellows of the AAAS. This past year the names of Albert Rees, Vernon Smith, and Sidney Winter were added to these ranks.

Further information about the AAAS is available from Marge White, 1333 H Street,

N.W., Washington, D.C. 20005. AEA members are also encouraged to contact members of the AEA/AAAS Liaison Committee: Roger Bolton, Gardner Brown, Faye Duchin, Adam Rose, Vernon Smith, and Robert Solow, about AAAS activities.

ADAM ROSE, *Representative*

Policy and Advisory Board of the Economics Institute

The Institute's first year under its new Director, Charles Becker, who succeeded Wyn Owen January 1, 1990, has been a good year. For several years prior to 1990, Economics Institute enrollments had declined because of developing country debt problems and a shortage of scholarship funds. These trends were reversed in 1990.

Total student enrollments increased almost one-third from 1989 to 1990, and student term equivalents increased 46 percent. Total revenues increased commensurately. Revenue increases enable the Institute to offer financial aid to more students whose funds cannot cover the full cost of the program. Enrollment increases were broadly based, with increasing numbers of students attending from many countries. Pervasive economic growth has enabled many countries to sponsor more students at the Institute and in U.S. graduate programs. In addition, U.S., source-country institutions, and international organizations have been able to sponsor more students.

Because of increased enrollments and revenues, the Institute has been able to broaden its preparatory work for graduate programs in economics, agricultural economics, and business. The Institute now offers a wide range of coursework in English, economic theory, mathematical economics, statistics, econometrics, and computer analysis. It is the only preparatory program that carefully integrates teaching of English with subject matter instruction. Especially impor-

tant has been expansion and upgrading of courses and laboratories in computing. Many students arrive at the Institute with little or no experience with modern computers. Computer facility is crucial not only for subsequent graduate study in the United States, but also for effective research and analysis after returning to home countries. Also important in upgrading the Institute's instructional program has been the acquisition of a state-of-the art Song language laboratory.

Among the large challenges that face the Institute in 1991 and subsequent years will be to satisfy the needs of substantial numbers of students from Eastern Europe and China.

Present Board members are: Daniel Bromley (University of Wisconsin-Madison), John Evans (University of North Carolina-Chapel Hill), Koichi Hamada (Yale University), Lawrence Lau (Stanford University), Ray Marshall (University of Texas-Austin), James Millar (University of Arkansas), Samuel Morley (Vanderbilt University), and Edwin Mills (Northwestern University).

Programs and institutions interested in further information about the Institute should contact Director, Economics Institute, 1030 13th Street, Boulder, CO 80302 (telephone 303-492-3000; Telex 450385 ECONINSTBDR).

EDWIN S. MILLS, *Chairman*

The American Economic Review

ARTICLES

- Popular Attitudes*
- ROBERT J. SHILLER, MAXIM BOYCKO, AND VLADIMIR KOROBV
Popular Attitudes Toward Free Markets: The Soviet Union
and the United States Compared
- MARTIN L. WEITZMAN
Price Distortion and Shortage Deformation, or
What Happened to the Soap?
- ALVIN E. ROTH
A Natural Experiment in the Organization of Entry-Level
Labor Markets: Regional Markets for New Physicians
and Surgeons in the United Kingdom
- SUSAN MONGELL AND ALVIN E. ROTH
Sorority Rush as a Two-Sided Matching Mechanism
- ATISH R. GHOSH AND PAUL R. MASSON
Model Uncertainty, Learning, and the Gains from Coordination
- SHLOMO YITZHAKI AND JOEL SLEMROD
Welfare Dominance: An Application to Commodity Taxation
- CHARLES W. CALOMIRIS AND CHARLES M. KAHN
The Role of Demandable Debt in Structuring Optimal
Banking Arrangements
- FRANK A. WOLAK AND CHARLES D. KOLSTAD
A Model of Homogeneous Input Demand Under Price Uncertainty
- LAURENCE BALL AND DAVID ROMER
Sticky Prices as Coordination Failure
- RICHARD E. ROMANO
When Excessive Consumption Is Rational
- EDWARD N. WOLFF
Capital Formation and Productivity Convergence Over the
Long Term
- GORDON LEITCH AND J. ERNEST TANNER
Economic Forecast Evaluation: Profits versus the
Conventional Error Measures
- SIMON BENNINGA AND ARIS PROTOPAPADAKIS
The Stock Market Premium, Production, and Relative
Risk Aversion
- DAVID DEJONG AND CHARLES H. WHITEMAN
The Temporal Stability of Dividends and Stock Prices:
Evidence from the Likelihood Function

SHORTER PAPERS: A. M. Polinsky and S. Shavell; G. E. Helfand; W. M. Hanemann;
T. Ellingsen; G. Gaudet and S. W. Salant; C. d'Aspremont, R. Dos Santos Ferreira, and
L.-A. Gérard-Varet; P. Rangazas; G. Edwards and D. Vanzetti.

JUNE 1991

THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

Officers

President

THOMAS C. SCHELLING
University of Maryland

President-elect

WILLIAM VICKREY
Columbia University

Vice-Presidents

HENRY J. AARON
The Brookings Institution
CLAUDIA D. GOLDIN
Harvard University

Secretary-Treasurer

C. ELTON HINSHAW
Vanderbilt University

Editor of The American Economic Review

ORLEY C. ASHENFELTER
Princeton University

Editor of The Journal of Economic Literature

JOHN PENCABEL
Stanford University

Editor of The Journal of Economic Perspectives

JOSEPH E. STIGLITZ
Stanford University

Executive Committee

Elected Members of the Executive Committee

STANLEY FISCHER
Massachusetts Institute of Technology
LAWRENCE H. SUMMERS
The World Bank
GREGORY C. CHOW
Princeton University
SUSAN ROSE-ACKERMAN
Yale University
MICHAEL J. PIORE
Massachusetts Institute of Technology
GAVIN WRIGHT
Stanford University

EX OFFICIO Member

GERARD DEBREU
University of California-Berkeley

• Printed at Banta Company, Menasha, Wisconsin.

• Copyright © American Economic Association 1991. All rights reserved.

• No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

Correspondence relating to advertising, business matters, permissions to quote, subscriptions, and changes of address, should be sent to the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Change of address notice must be received at least six (6) weeks prior to the publication month. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

THE AMERICAN ECONOMIC REVIEW (ISSN 0002-8282), June 1991, Vol. 81, No. 3, is published five times a year (March, May, June, September, December) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Annual fees for regular membership, of which 30 percent is for a year's subscription to this journal, are: \$44.00, \$52.80, or \$61.60 depending on income. A membership also includes the *Journal of Economic Literature* and the *Journal of Economic Perspectives*. In countries other than the U.S.A., add \$16.00 for extra postage. Second-class postage paid at Nashville, TN and at additional mailing offices. POSTMASTER: Send address changes to the *American Economic Review*, 2014 Broadway, Suite 305, Nashville, TN 37203.

VICTOR R. FUCHS

DISTINGUISHED FELLOW

1990

Victor R. Fuchs is widely recognized as one of the founders of health economics, and his research has touched on every aspect of this field. He also has done important research in industrial organization and labor economics. These latter contributions are highlighted in his fascinating work on the service economy. Here one finds a variety of carefully crafted empirical studies on trends in productivity and employment in the service, industry, and agricultural sectors; inter-industry and intersectoral differences in hourly earnings; and cyclical fluctuations in productivity.

His research in health consists of studies ranging from the economic determinants of mortality rates and the roles of schooling, time preference, and cigarette smoking in health outcomes, to detailed analyses of the markets for physicians' and surgeons' services and to the implications of health policy reforms in the United States. This research reflects themes for which Fuchs has become justly famous, including the importance of individual behavior in health outcomes and the key role played by physicians in the market for their services.

Fuchs has stressed that research should be empirically sound and policy relevant. He also is to be commended for his prodigious efforts to make the key insights of economics available to physicians and policymakers. The American Economics Association honors a truly outstanding empirical researcher who uses economic theory to better understand and improve the environment in which we live.



Victor R. Zuck

THE AMERICAN ECONOMIC REVIEW

Editor

ORLEY ASHENFELTER

Co-Editors

ROBERT H. HAVEMAN
BENNETT T. McCALLUM
PAUL R. MILGROM

Acting Co-Editor

JOHN Y. CAMPBELL

Production Editor

J. DAVID BALDWIN

Board of Editors

JOSEPH G. ALTONJI
JAMES E. ANDERSON
ALAN J. AUERBACH
KYLE W. BAGWELL
ROBIN W. BOADWAY
TIMOTHY F. BRESNAHAN
LORNE H. CARMICHAEL
GEORGE W. EVANS
HENRY S. FARBER
MARJORIE A. FLAVIN
ROBERT P. FLOOD
JO ANNA GRAY
REUBEN GRONAU
DANIEL S. HAMERMESH
ROBERT J. HODRICK
KEVIN D. HOOVER
JOHN H. KAGEL
JOHN F. KENNAN
JOHN McMILLAN
JOHN D. ROBERTS
RICHARD ROLL
THOMAS ROMER
DAVID E. M. SAPPINGTON
SUZANNE A. SCOTCHMER
HAL R. VARIAN
KENNETH WEST
JOHN D. WILSON
LESLIE YOUNG

June 1991

VOLUME 81, NUMBER 3

Articles

- Popular Attitudes Toward Free Markets: The Soviet Union and the United States Compared
Robert J. Shiller, Maxim Boycko, and Vladimir Korobov 385
- Price Distortion and Shortage Deformation, or What Happened to the Soap?
Martin L. Weitzman 401
- A Natural Experiment in the Organization of Entry-Level Labor Markets: Regional Markets for New Physicians and Surgeons in the United Kingdom
Alvin E. Roth 415
- Sorority Rush as a Two-Sided Matching Mechanism
Susan Mongell and Alvin E. Roth 441
- Model Uncertainty, Learning, and the Gains from Coordination
Atish R. Ghosh and Paul R. Masson 465
- Welfare Dominance: An Application to Commodity Taxation
Shlomo Yitzhaki and Joel Slemrod 480
- The Role of Demandable Debt in Structuring Optimal Banking Arrangements
Charles W. Calomiris and Charles M. Kahn 497
- A Model of Homogeneous Input Demand Under Price Uncertainty
Frank A. Wolak and Charles D. Kolstad 514
- Sticky Prices as Coordination Failure
Laurence Ball and David Romer 539
- When Excessive Consumption Is Rational
Richard E. Romano 553
- Capital Formation and Productivity Convergence Over the Long Term
Edward N. Wolff 565
- Economic Forecast Evaluation: Profits versus the Conventional Error Measures
Gordon Leitch and J. Ernest Tanner 580
- The Stock Market Premium, Production, and Relative Risk Aversion
Simon Benninga and Aris Protopapadakis 591
- The Temporal Stability of Dividends and Stock Prices: Evidence from the Likelihood Function
David N. DeJong and Charles H. Whiteman 600

Shorter Papers

A Note on Optimal Fines When Wealth Varies Among Individuals	<i>A. Mitchell Polinsky and Steven Shavell</i>	618
Standards versus Standards: The Effects of Different Pollution Restrictions	<i>Gloria E. Helfand</i>	622
Willingness to Pay and Willingness to Accept: How Much Can They Differ?	<i>W. Michael Hanemann</i>	635
Strategic Buyers and the Social Cost of Monopoly	<i>Tore Ellingsen</i>	648
Increasing the Profits of a Subject of Firms in Oligopoly Models with Strategic Substitutes	<i>Gérard Gaudet and Stephen W. Salant</i>	658
Pricing-Schemes and Cournotian Equilibria	<i>Claude d'Aspremont, R. Dos Santos Ferreira, and L.-A. Gérard-Varet</i>	666
Redistribution and Capital Formation	<i>Peter Rangazas</i>	674
The Welfare Economics of Price Supports in U.S. Agriculture: Comment	<i>Geoff Edwards and David Vanzetti</i>	683
Auditors' Report		685

•Submit manuscripts (4 copies); 50 pages maximum, single-sided, double-spaced, to:

Orley Ashenfelter, Editor, *AER*; 209 Nassau Street, Princeton, NJ 08542-4607.

•Submission fee: \$50 for members; \$100 for nonmembers. Please pay with a check or money order payable in United States Dollars. Canadian and Foreign payments must be in the form of a check drawn on a United States bank payable in United States Dollars. Style guides will be provided upon request.

Editorial Statement

It is the policy of the *American Economic Review* to publish papers only where the data used in the analysis are clearly and precisely documented, are readily available to any researcher for purposes of replication, and where details of the computations sufficient to permit replication are provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary, or if, for some other reason, the above requirements cannot be met.

Popular Attitudes Toward Free Markets: The Soviet Union and the United States Compared

By ROBERT J. SHILLER, MAXIM BOYCKO, AND VLADIMIR KOROBOV*

Random samples of the Moscow and New York populations were compared in their attitudes toward free markets by administering identical telephone interviews in the two countries in May 1990. Although the Soviet respondents were somewhat less likely to accept exchange of money as a solution to personal problems and although their attitudes toward business were less warm, we found that the Soviet and American respondents were basically similar in some very important dimensions: in their attitudes toward fairness, income inequality, and incentives and in their understanding of the working of markets. (JEL O57)

What are the important barriers to the success of free markets? At this time of transition in the Soviet Union and other Eastern economies, the answer to this question is of the utmost importance. One view is that major obstacles are the attitudes, morals, and understandings of the people themselves, not just the institutions or politics they live with. Leonid Abalkin, former Deputy Prime Minister of the Soviet Union and prominent economist, complained that

...it is not easy to develop a stratum of talented people, with a good under-

standing of the market. For that, it is necessary to put aside fixed patterns of thinking, inherited from the past, to consider afresh our morals, and our system of values in general.

(Abalkin, 1990 p. 9)

This has been a recurring theme, appearing quite often in the Soviet Parliament and government bodies, in the mass media, and in academic journals: it is argued that the general public in the Soviet Union is not prepared to accept and fully use markets.

For an accurate study of the validity of this argument for the Soviet Union, one must have an explicit comparison group (another country) and compare carefully between the countries the frequencies of public understandings, values, attitudes, and behaviors relating to markets. People everywhere are, to some extent or with some frequency, resistant to market solutions, and when the differences in frequencies across countries are not total the casual observer of people cannot be expected to keep an accurate count. The importance of having a comparison or control group in research is of course well known, and the history of science shows many examples of carefully controlled studies overturning formerly "unassailable" theories.

We have undertaken surveys of randomly selected individuals in the Soviet Union and

*Shiller is the Stanley B. Resor Professor of Economics, Yale University, New Haven, CT 06520; Boycko is an economist at the Institute of World Economy and International Relations, U.S.S.R. Academy of Sciences, Profsoyuznaya 23, Moscow 117869; Korobov is a sociologist at the Institute for Sociology, U.S.S.R. Academy of Sciences, Krzizanovskogo 24/35 b. 5, Moscow 117259. The authors are grateful to Robert Abelson, Sergei Aukuzionek, Revold Entov, Martin Feldstein, Daniel Kahneman, Alvin Klevorick, William Mahota, Michael Montias, Andrey Poletaev, Thomas Richardson, Lawrence Summers, Richard Thaler, and participants at a seminar at the Cowles Foundation at Yale University for helpful comments. This paper was supported by the Institute of World Economy and International Relations of the U.S.S.R. Academy of Sciences, the Institute for Sociology of the U.S.S.R. Academy of Sciences, the Russell B. Sage Foundation, and the U.S. National Science Foundation. Any opinions are those of the authors and not necessarily those of their respective institutions.

in the United States with questions aimed at finding out about opinions concerning whether price changes are fair, attitudes toward income inequality, popular theories concerning the importance of incentives, inhibitions against exchange of money, envy or hostility toward business people and the rich, popular understandings of markets and speculation, understandings of the welfare effects of compensated price changes, and expectations of future government interference.

We use questions that are aimed at providing evidence on *fundamental* parameters of human behavior related to the success of free markets. Sometimes our questions are about aspects of everyday life that are not directly affected by government economic policies. Other questions are about basic economic intuition. Sociologists have noted that popular answers to such questions often differ substantially from the answers that would be suggested by the dominant ideology that is expressed by opinion leaders.¹

To our knowledge, there has never been a similar study that examines attitudes relevant to the functioning of free markets in these countries. There have, of course, been many recent opinion surveys in the Soviet Union, and a few of these ask questions that are relevant to the success of free markets. For example, one study (Yu. A. Levada, 1990) asked Soviets what kinds of changes they would consider to be important indicators of great improvement in the situation of the country. The authors of the study concluded that respondents tended to give high rank to general concepts like "bringing goods back on the counter" or stable prices and that "values that are asso-

ciated with economic and civil freedom have much fewer supporters" (Levada, 1990 p. 50).² The major surveys of attitudes of Soviets published in the United States, based on responses of Soviet emigres, were not directly interested in the potential for success of free markets; however, they did produce a few results tangentially relevant here. The Harvard Refugee Interview Project³ concluded from surveys in 1950-1951 that Soviet emigres tended to support government control of the economy, expressed a strong dedication to society over the individual, and reacted negatively to Western materialism. The Soviet Interview Project⁴ concluded from surveys in the early 1980's that Soviet emigres tended to support state control of medicine but not of agriculture and that most supported the right of workers to strike. However, there does not appear to be a lot more in these studies on attitudes toward free markets. No Western or Soviet study relevant to free markets that we know of has made explicit comparisons of Soviets with Westerners.

I. Questionnaire Design and Survey Methods

Our questionnaires included 36 questions, addressing various aspects of human behavior related to free markets.⁵ Some of our questions probed public opinion on certain issues, but mostly the respondents were asked to consider some imaginary situation that they might experience and to describe their behavior in, or judgment of, that situation.

Naturally, when evaluating responses there is always some doubt whether they

¹Nicholas Abercrombie et al. (1980 p. 141), after reviewing a variety of interview results, asserted that people "will often agree with dominant elements, especially when these are couched as abstract principles or refer to general situations, which is normally the case in interview surveys using standardised questionnaires, but will then accept deviant values when they themselves are directly involved or when these are expressed in concrete terms which correspond to everyday reality."

²For a summary of other relevant Soviet public-opinion research, see Hans Aage (1991).

³For an overview of the results of this project, see Raymond Bauer et al. (1957).

⁴For an overview of this project, see James Millar (1987).

⁵Original questionnaires in both English and Russian, as well as further information about the samples and statistical methods, are available from the Cowles Foundation, Yale University, as part of Discussion Paper No. 952, August 1990.

were really determined by the basic attitudes we are interested in and not by the specifics of a particular scenario. To develop confidence in our results, we usually asked a number of similar questions placed in different contexts (and sometimes even addressed to different subsamples). When there are similar responses to these questions, we feel that we have some grounds to generalize beyond the specifics of the particular situations. In a sense, it is the totality of all the questions asked that gives us more confidence in the results reported below.

Still, we think that the evidence is mostly suggestive, not assertive. In some cases, the results just indicate that certain types of beliefs about the Soviets and Americans are at odds with the evidence that we have. Although we do not claim to settle the issues here, we think that our results do provide some substantial evidence.

When designing the questions, we tried to do our best to make them equally comprehensible to the Soviet and the American respondents. For that, first of all, we took great care in selecting our scenarios of imaginary situations that would possibly make the same sense for both audiences, despite the very different institutional environments that they generally face. For instance, one of our questions (B2) described a price increase at a flower market due to soaring demand on the eve of a holiday. This is a rare instance of a temporary price increase that the Soviets are quite familiar with. Similarly, when comparing price and nonprice rationing methods (question C4), we used gasoline as our example because Americans may still remember President Carter's standby gasoline rationing plan of 1979, or the odd-day/even-day gasoline rationing scheme actually imposed by some eastern states then.

Second, we put a lot of effort into selecting suitable wordings, so that the questions would sound as much alike as possible in the two languages. Originally the questionnaire was developed in English, but then we made several rounds of translating it into Russian and back, each time adjusting the wordings where appropriate. We also usually said something like "5 percent" rather

than "a little" to reduce further ambiguities in translation.⁶

The survey was conducted by means of telephone interviews with randomly selected individuals of 18 years of age or older. We documented responses from 391 residents of Moscow and 361 residents of the greater New York City consolidated metropolitan statistical area. The 36 questions were subdivided into three parts (designated A, B, and C in the question numbers), and each respondent was asked to answer only one part consisting of 12 questions. We were able to document about 120–130 responses per question in each country. The two samples were generally representative of their underlying populations and also rather close to each other in terms of basic characteristics: average age was 45 in the Soviet Union and 42 in the United States; 60 percent of U.S.S.R. respondents and 58 percent of U.S. respondents were female; 50 percent of U.S.S.R. respondents and 66 percent of U.S. respondents had attended some college. In both countries, those who agreed to participate in the survey were perhaps a little more articulate and informed than average for their respective populations, but it is our impression that as a result they generally had no difficulty understanding the questions.

The closeness of characteristics of the samples makes it generally possible to attribute any differences that we find to genuine differences between Soviets and Americans and not to differences in the composition of our samples.⁷ However, we

⁶In an independent evaluation of our translation, William Mahota, Professor of Slavic Languages and Literature, Yale University, wrote, "I have closely compared the Russian and English versions of Shiller, Boycko, and Korobov's survey of attitudes toward economic problems, and found that the language of the two versions corresponds virtually exactly." (See the earlier version of this paper: Shiller et al., 1990.)

⁷With sample sizes of a little over 100, the standard error of an estimated proportion is just under 5 percent if the estimated proportion is 50 percent; it is 4 percent if the estimated proportion is 25 percent or 75 percent; and it is 3 percent if the sample estimated proportion is 10 percent or 90 percent. Thus, for exam-

have also carried out probit regressions that allow us to evaluate the statistical significance of the intercountry differences when other observable characteristics are controlled for. When presenting our results below, we report *t* statistics of the coefficient of the country dummy variable in a probit regression. All estimated equations for each question had a constant term and the same standard set of right-hand-side predictors: namely, dummies for country, sex, and rural origin and also respondent's age and education level (based on an index from 1 to 6, with 1 representing "did not finish high school" and 6 indicating "finished graduate school").

There are somewhat fewer telephones per household in Moscow. At the end of 1988, there were 2.70 million telephones in private apartments in Moscow; at the same time, there were 3.05 million private apartments, implying 89 telephones per 100 apartments (*Moskva v Tsifrach*, 1989). In 1990, an estimated 93 percent of all households in the New York consolidated metropolitan statistical area had telephones. Only 61 percent of New York households had listed phones,⁸ but with random-digit dialing, nonlisting does not affect results.

An obvious criticism of our samples is that Moscow is probably not representative of the Soviet Union at large; the people there may be a little more educated or aware of economic issues. However, New York City, sometimes referred to as the business and financial "capital" of the United States, may also be populated by those who are more "advanced" in their attitudes toward markets than the rest of the country, so that the intracountry bias is possibly in the same direction. Even if this argument is not entirely convincing, we feel that a comparison between the two cities is quite meaningful by itself. The respondents in our two samples may represent the more

economically active and influential people in the two countries. Thus, our results may be more relevant to understanding economic events in the two countries than if we had taken a representative sample of everyone in the two countries.

II. Fairness of Price Changes

One important potential obstacle to the clearing of free markets is a popular feeling that price increases may be unfair. If sellers feel that they cannot raise their prices, then they will be forced to use nonprice rationing to distribute their goods, contrary to market principles.

It is widely believed in the Soviet Union (and possibly elsewhere) that the Soviet people, being for a long time accustomed to stable, government-sanctioned prices, will be characteristically reluctant to accept market prices. Consider the following statement of V. O. Rukavishnikov, a prominent Soviet sociologist:

...[T]he public attitude towards possible increases of prices of consumer goods that are in short supply is extremely negative, because this solution to the problem of the queues is likely to lead to a situation with lots of goods on the counters, with no queues, but with few people being able to buy the goods; 83.7 percent of the people surveyed are against this solution. [4.4 percent support it, and 11.9 percent did not answer.]⁹

(Rukavishnikov, 1989 p. 4)

Such a result may reflect general human behavior, not just Soviet behavior. In North American survey results, Daniel Kahneman et al. (1986) have also documented much resistance to price increases that were considered unfair.

For a meaningful evaluation of the attitudes toward free prices in the Soviet Union, it is useful to compare Soviets and Americans responding to identical questions in

ple, an estimated sample proportion of 25 percent has a 95-percent confidence interval of 17-33 percent.

⁸The U.S. data were provided courtesy of Survey Sampling, Inc., Fairfield, CT.

⁹The figures are based on about 5,000 responses sent to the popular magazine *Sobesednik* by its readers in September and October 1988.

identical contexts. We report several similar scenarios (inspired by Kahneman et al.), designed to address this issue:

B2. *On a holiday, when there is a great demand for flowers, their prices usually go up. Is it fair for flower sellers to raise their prices like this?*

Response	U.S.S.R.	U.S.A.	t [1 vs. 2] ¹⁰ (d.f.)
1) Yes	34%	32%	-0.89
2) No	66%	68%	(241)
	N: 131	119	

B11. *A small factory produces kitchen tables and sells them at \$200 each. There is so much demand for the tables that it cannot meet it fully. The factory decides to raise the price of its tables by \$20, when there was no change in the costs of producing tables. Is this fair?*

Response	U.S.S.R.	U.S.A.	t [1 vs. 2] (d.f.)
1) Yes	34%	30%	-0.71
2) No	66%	70%	(242)
	N: 131	120	

A9. *A new railway line makes travel between city and summer homes positioned along this rail line substantially easier. Accordingly, summer homes along this railway become more desirable. Is it fair if rents are raised on summer homes there?*

Response	U.S.S.R.	U.S.A.	t [1 vs. 2] (d.f.)
1) Yes	57%	61%	0.06
2) No	43%	39%	(199)
	N: 98	115	

The critical point here is that there is virtually no difference between U.S.S.R. and U.S. answers. In the first two scenarios, we

discover a tendency in *both* countries to report that price increases are unfair. In the third scenario,¹¹ in *both* countries most people think that price increases are fair. Here, our comparison-group methodology displays its full power. Notions of fairness are very situation-specific: flower sellers are unfair if they raise their prices, while landlords who do so in the circumstance described are usually not. Notions of fairness are *not* country-specific. The bottom line from all of this is that there is little foundation to the aforementioned claims that Soviets are *characteristically* resistant to unfair price increases.

We were able to expand our understanding of fairness by asking about the policy implications of the fairness judgments. After the question about flower sellers we asked:

B3. *Should the government introduce limits on the increase in prices of flowers, even if it might produce a shortage of flowers?*

Response	U.S.S.R.	U.S.A.	t [1 vs. 2] (d.f.)
1) Yes	54%	28%	-3.71
2) No	46%	72%	(229)
	N: 123	115	

After the question about the manufacturer of tables we asked:

B12. *Apart from fairness, should the factory have the **right** to raise the price in this situation?*

Response	U.S.S.R.	U.S.A.	t [1 vs. 2] (d.f.)
1) Yes	57%	59%	0.29
2) No	43%	41%	(227)
	N: 118	118	

¹⁰Throughout this article, the t statistic is from a probit regression, as described above. In brackets, we indicate the construction of the binary choice variable.

¹¹Perhaps more striking than the majority who think a rent increase is fair is that Americans were more ready to provide a definite opinion; the response rates were as follows: U.S.S.R. = 75 percent; U.S.A. = 96 percent. This kind of difference was encountered rather often in our results, but it is of secondary importance for the purposes of this study.

In only one of these two questions, the first (B3), was there a significant difference between Soviet and American responses. Soviets are more likely to want to restrict the flower seller from raising prices, but both Soviets and Americans tend to agree that the manufacturer of tables has, in effect, the right to be "unfair." (The answers to the second question (B12) show that, in both countries, beliefs that something is unfair need not translate into an opinion that something should be illegal.)

Another perspective on the fairness issue can be gained by posing a question without the word "fair," but asking whether an action is "moral." Here, we have changed the context of the question to a price increase between sale and resale, raising the issue of profiteering:

C10. *A small merchant company buys vegetables from some rural people, brings the vegetables to the city, and sells them, making from this a large profit. The company honestly and openly tells the rural people what it is doing, and these people freely sell the company the vegetables at the agreed price. Is this behavior of the company, making large profits using the rural people, acceptable from a moral point of view?*

Response	U.S.S.R.	U.S.A.	<i>t</i> [1 vs. 2] (d.f.)
1) Yes	49%	59%	0.52
2) No	51%	41%	(218)
	N: 120	116	

Again, the Soviets are not dramatically more concerned with profiteering, and this difference is not statistically significant.

We wanted to learn whether people would impose on themselves the hardships caused by rationing of quantities, and so we asked:

C4. *Suppose that the government wishes to reduce consumption of gasoline. They propose two methods of attaining this goal. First, the government could prohibit gas stations from selling, for example, more than five gallons to one person. Second, the government could put a tax on gasoline, and prices of gasoline*

would go up. From your point of view, which of these methods is better?

Response	U.S.S.R.	U.S.A.	<i>t</i> [1 vs. 2] (d.f.)
1) First	43%	36%	-1.28
2) Second	57%	64%	(196)
	N: 104	109	

Now, neither the Soviets nor the Americans tend to think that it is a good idea to force people to buy gasoline in small quantities. The Americans were only slightly less likely to favor the rationing solution.

Overall, the reported evidence suggests that there is actually little ground to believe that the Soviets are characteristically more hostile toward free-market prices. The strong opposition to price reform (implying price increases) that undoubtedly exists in the Soviet Union should not be attributed to peculiarities of national character; rather, the economic and political interests should be given more weight. Obviously, by setting prices free, central planners will lose an important instrument of control over the enterprises as well as some arbitrage opportunities that result from disequilibrium pricing. (For additional evidence on attitudes toward price changes, see the response to question B6 in Section VII and questions C6 and B10 in Section VIII.)

III. Attitudes Toward Income Inequality

Popular notions of fairness are essentially related to attitudes toward inequality. Given the history of Communist ideology, it would seem that Soviet citizens might be more intolerant of inequalities of income and wealth. Of course, "from each according to his abilities, to each according to his needs" has long been a Communist slogan. With the reputation of the United States as the most capitalist country, it would seem that American citizens might be much more tolerant of inequalities of income and wealth. However, we found no evidence to support such a notion.

One question, designed to see whether people would object to pro-market reforms

because of envy of those people who would succeed under such reforms, found that the *Americans* were the most resistant:

A4. *Suppose the government wants to undertake a reform to improve the productivity of the economy. As a result, everyone will be better off, but the improvement in life will not affect people equally. A million people (people who respond energetically to the incentives in the plan and people with certain skills) will see their incomes triple while everyone else will see only a tiny income increase, about 1 percent. Would you support the plan?*

Response	U.S.S.R.	U.S.A.	<i>t</i> [1 vs. 2] (<i>d.f.</i>)
1) Yes	55%	38%	-2.07
2) No	45%	64%	(199)
	N: 114	99	

The plan described makes everyone better off, so any objections would have to be motivated by the relative inequality created by the plan. Only about half of the Soviet respondents supported the plan, but even fewer of the U.S. respondents responded that way.

Another way to quantify attitudes toward income inequality is to ask respondents about how they would tax inheritances of the rich:

A10. *In your opinion, what inheritance tax rate for really wealthy people do you think we should have? A tax rate of 0 percent means that they can pass all of their wealth to their children, making them as rich as their parents. A rate of 50 percent means that they can pass half to their children. A rate of 100 percent means that they can pass none at all to their children.*

Rate	U.S.S.R.	U.S.A.
Mean	39%	37%
Median	34%	30%
	N: 99	107

There was virtually no difference between the Soviet and American answers.

IV. Popular Theories about the Importance of Incentives

One theory to explain the slowness of the Soviet Union to implement a market system is that people there do not believe in one of its alleged principal advantages: the incentives that the system creates for hard work. The Soviets are reputed not to think that most people are basically motivated for personal gain and to believe instead that people work better if they are in a social context that makes their work personally meaningful to them.

When our respondents were asked directly about this, it turned out that there was very little difference between the Soviet and American responses.¹²

A1. *Do you think that people work better if their pay is directly tied to the quantity and quality of their work?*

Response	U.S.S.R.	U.S.A.	<i>t</i> [1 vs. 2] (<i>d.f.</i>)
1) Yes	90%	86%	-1.05
2) No	10%	14%	(226)
	N: 121	119	

We asked much the same question in a different way, in terms of the important qualities of managers:

C3. *Which of the following qualities is more important for the manager of a company?* [Response choices: 1) *The manager must show good will in his relation to workers and win their friendship;* 2) *The manager must be a strict enforcer of work discipline, giving*

¹²An earlier survey (Tom W. Smith, 1989) allows comparisons among seven different countries, including one socialist country (Hungary), in respondent's answers to a similar question. Respondents were asked how much they agreed with the statement "financial incentives are needed if people are to work hard." Of U.S. respondents, 68 percent agreed or strongly agreed; of Hungarian respondents 70 percent agreed or strongly agreed. In another survey in the Soviet Union (Yu. A. Levada, 1990) only 40 percent of a nationally representative urban sample agreed with the strongly worded statement "without big differences in labor remuneration there will be no incentives to good work."

incentives to hard workers and punishing laggards.]

Response	U.S.S.R.	U.S.A.	$t[1 \text{ vs. } 2]$ (d.f.)
1	33%	49%	2.65 (204)
2	68%	51%	
	N: 112	109	

Again, it is the Soviets, not the Americans, who tend to believe in strict managers.

We also asked our respondents if they had heard about the capitalist theory that, because of the importance of incentives, income inequality is a necessary evil:

A2. Some have expressed the following: "It's too bad that some people are poor while others are rich. But we can't fix that: if the government were to make sure that everyone had the same income, we would all be poor, since no one would have any material incentive to work hard." Have you heard such a theory or not? If yes, then how often?

Response	U.S.S.R.	U.S.A.	$t[(1+2) \text{ vs. } 3]$ (d.f.)
1) Often	38%	7%	-4.89
2) Once or twice	39%	38%	(231)
3) Never heard it	23%	55%	
	N: 125	120	

Surprisingly, the Soviet respondents were more familiar with this theory than their U.S. counterparts, perhaps due to current extensive discussions of this and related subjects in the Soviet mass media.

A3. Do you yourself personally agree with this theory?

Response	U.S.S.R.	U.S.A.	$t[1 \text{ vs. } 2]$ (d.f.)
1) Yes	41%	38%	-0.48
2) No	59%	62%	(213)
	N: 110	116	

Neither country seems to like this theory a lot, but the opposition to the theory is weaker among our respondents in the Soviet Union. It is the American responses that are more surprising here. Agreement with this theory is not actually contrary to Communist theory of the past 20 or so years. Alistair McAuley (1980 p. 242), in a survey of Soviet academic economists and lawyers, concludes that "most Soviet economists appear to advocate what one might call a meritocratic structure of wages."¹³

V. Resistance to Exchange of Money

The essence of a market system is the ability of persons to secure the things they want by the voluntary and unrestricted exchange of money. Such "creative" exchanges of money are quite different from the exchanges of money that might be sanctioned by a government agency that certifies that the transaction is fair and equitable. We hypothesized that considerations of fairness, equity, and friendship might inhibit such exchanges relatively more in the Soviet Union.

The charging of interest to others for a loan is a practice that has been censured as immoral since ancient times, but of course certain forms of interest payments have legal sanction in both the Soviet Union and the United States today. We sought to abstract from the current legal environment by describing a hypothetical situation between friends:

A7. Suppose you have agreed to lend a friend some money for six months, so that he will not miss a good opportunity to buy a summer home. Suppose banks are offering interest rates of 3 percent per year. Would you charge him interest on the loan?

¹³In Smith's survey, it was found that 25 percent of Hungarian and 31 percent of U.S. respondents either agreed or strongly agreed with the statement "large differences in income are necessary for national prosperity (Smith, 1989 p. 70).

Response	U.S.S.R.	U.S.A.	<i>t</i> [1 vs. 2] (d.f.)
1) Yes	6%	29%	4.27
2) No	94%	71%	(215)
	N: 117	111	

The difference here is quite substantial: about five times as many U.S. respondents answered "yes," compared to their Soviet counterparts. Although most people in both countries said that they would not charge a friend interest, we interpret these results as implying that there is a much bigger minority in the United States who are accustomed to an exchange of money as a solution to everyday problems.

Still, it is not entirely clear that the difference reported is truly attitudinal, and not institutional. Even though the question specifies the rate of interest at 3 percent, U.S. respondents are more familiar with high interest rates and may therefore have learned in the past that lending money to a friend at zero interest can be costly. We sought, therefore, to find a question that is relatively unrelated to past market experience. We asked:

A8. *If you went on a vacation with friends and there were a lot of shared expenses, would there be a careful accounting of who spent what and a settling of accounts afterwards?*

Response	U.S.S.R.	U.S.A.	<i>t</i> [1 vs. 2] (d.f.)
1) Yes	30%	47%	2.66
2) No	70%	53%	(221)
	N: 116	118	

Here again is some evidence that U.S. respondents are rather more accustomed to an exchange of money, although the difference is less striking than with the previous question about charging interest.

Another question that would appear to abstract from any different experience with market solutions in the situation described is the following:

B7. *You are standing in a long line to buy*

something. You see that someone comes to the line and is very distressed that the line is so long, saying he is in a great hurry and absolutely must make this purchase. A person at the front of the line offers to let him take his place in line for \$10.00. Would you be annoyed at this deal even though it won't cause you to wait any longer?

Response	U.S.S.R.	U.S.A.	<i>t</i> [1 vs. 2] (d.f.)
1) Yes	69%	44%	-3.61
2) No	31%	56%	(240)
	N: 132	117	

Clearly, the Soviet respondents showed substantially more annoyance at the deal described than did the Americans. This annoyance is noteworthy, since the deal apparently is helping a distressed person and since the deal harms no one else. Such annoyance at harmless interruptions in line has been noted before (see Jon Elster, 1989).

As before, the difference in responses may be attributed to the specifics of economic conditions in the two countries. Currently, the queues constitute a major concern for the Soviet consumer, and he has ample reason to be quite touchy in this respect.

When evaluating responses to all three questions in this section taken together, a common explanation looks at least as persuasive as several specific ones. Therefore, we conclude that there is some evidence that the Soviets are, to a certain extent, less willing to accept exchange of money as a solution to their problems. (For supplementary evidence on these issues, see questions C1, B4, and A5 in Section VI.)

VI. Negative Attitudes Toward Business

Many scholars have claimed that the Russian people have a long-standing aversion to business and dislike of businessmen. Alexander Gerschenkron (1962 p. 60) wrote that "There is no doubt that throughout most of the nineteenth century a grave opprobrium attached to the entrepreneurial

activities in Russia Divorced from the peasantry, the entrepreneurs remained despised by the intelligentsia." The idea is commonplace that the Communist revolution may have had its roots partly in such feelings. We sought to find whether there is evidence that such feelings today really set Soviet citizens apart from their U.S. counterparts.

We sought first to find whether people in the two countries feel that they would be esteemed by their relatives and friends if they were successful in business:

C1. Suppose that as a result of successful business dealings you unexpectedly became rich. How do you imagine it would be received by your relatives at a holiday family gathering? Would they congratulate you and show great interest, or would they be judgmental and contemptuous? [Response choices: 1) They would show interest, would congratulate; 2) They would be judgmental and contemptuous; 3) They would be quiet, indifferent.]

Response	U.S.S.R.	U.S.A.	t[1 vs. (2+3)] (d.f.)
1	72%	92%	2.08
2	12%	6%	(194)
3	16%	3%	
N: 113		117	

The Americans get greater support from their relatives and friends, though most of the Soviets expect congratulations.

C9. Do you think that, if you worked independently today as a businessman and received profit, your friends and acquaintances would respect you less and not treat you as you deserve?

Response	U.S.S.R.	U.S.A.	t[1 vs. 2] (d.f.)
1) Yes	19%	4%	-3.04
2) No	81%	96%	(216)
N: 115		120	

This evidence suggests that on the whole

neither country lacks respect for businessmen, but there is less respect for them in the Soviet Union.

A somewhat different attitude toward business that we wished to explore is whether people relish the prospect of showing off their wealth, and whether or not that helps them find good friends:

C2. If you ever became rich, would you really like to spend some of the money by purchasing really fashionable clothes, expensive cars, or other extravagant items that make an impression on people?

Response	U.S.S.R.	U.S.A.	t[1 vs. 2] (d.f.)
1) Yes	35%	50%	1.60
2) No	65%	50%	(217)
N: 115		120	

These responses may be interpreted as indicating that the Americans find the life of a successful businessman more appealing or want to show off a bit more. However, this may just be the result of their better circumstances. Levels of aspiration are affected by the standards of one's peers and are raised by a sense of accomplishment or success; this point has been stressed by psychologists (see Kurt Lewin et al., 1944). One may argue also that the Soviets, with a substantially lower standard of living, simply have more immediate concerns on their minds than thinking about what to do when they become rich. They may not have found it worthwhile to expend the costs of collecting information about luxury goods.

A way of getting at attitudes toward success in business without mentioning specific purchases is to make people choose between a general notion of success in business or in some other arena of life:

B4. Which of the following achievements would please you more? [Response choices: 1) You win fortune without fame; you make enough money through successful business dealings so that you can live very comfortably

for the rest of your life; 2) You win fame without fortune: for example you win a medal at the Olympics or you become a respected journalist or scholar.]

Response	U.S.S.R.	U.S.A.	$t[1 \text{ vs. } 2]$ (d.f.)
1	65%	54%	-1.47 (201)
2	35%	46%	
N: 92		117	

Although the U.S. respondents answered the question much more freely (response rates: U.S.S.R. = 67 percent; U.S.A. = 98 percent), of those who did answer the Soviets were relatively more attracted by wealth.

A5. *Is it important to you that your work benefits the country, and is not just to make money? Is it very important, somewhat important, or not important?* [Response choices: 1) Very important; 2) Somewhat important; 3) Not important.]

Response	U.S.S.R.	U.S.A.	$t[(1+2) \text{ vs. } 3]$ (d.f.)
1	69%	40%	-2.25
2	25%	45%	(235)
3	6%	15%	
N: 130		119	

The U.S. respondents are more for the money here, though of course we could also interpret the results as indicating that they feel freer to *admit* this.¹⁴

Yet another way to get at attitudes toward business success is to try to elicit from respondents their prejudices against businessmen:

C11. *Do you think that it is likely to be difficult to make friends with people who have*

their own business (individual or small corporation) and are trying to make a profit?

Response	U.S.S.R.	U.S.A.	$t[1 \text{ vs. } 2]$ (d.f.)
1) Yes	51%	20%	-4.65
2) No	50%	80%	(214)
N: 111		121	

On this question, Soviets are much less sanguine about businessmen than are the Americans.

C5. *Do you think that those who try to make a lot of money will often turn out to be not very honest people?*

Response	U.S.S.R.	U.S.A.	$t[1 \text{ vs. } 2]$ (d.f.)
1) Yes	59%	39%	-2.23
2) No	41%	62%	(214)
N: 114		117	

Indeed, relatively more Soviets do tend to expect businessmen to be less honest.

These last two questions show that U.S.S.R. respondents attach negative prejudices toward businessmen; but a caveat is in order. When evaluating these prejudices, it is important to keep in mind that many Soviets have never met a businessman in an informal situation, to say nothing of knowing one well. Their answers may be determined by what they read or hear, not by personal experience.

Still, the prejudices that Soviets have today are probably obstacles toward development of business enterprises. The questions in this section, which have various interpretations individually, tend generally to support the notion that Soviets indeed display a somewhat less warm attitude toward business and may be less interested in business careers.

It should of course be borne in mind that the differences we found were often value differences, differences in what each person

¹⁴Bauer et al. (1957 p. 128) concluded from the Harvard Refugee Interview Project that Soviet emigres felt that "we [Westerners] are lacking in spiritual and cultural values, in altruism, in dedication to society."

wants in his or her own life. Perhaps economists should not argue over them or be concerned about them.

VII. Perceptions of Speculation

Many barriers to free market activity are supported in the Soviet Union on the ground that these activities represent "speculation." Unfortunately, the term "speculation" has a wide range of meanings. Sometimes the term "speculation" in the Soviet Union refers to activities that consist of taking (in effect stealing) goods intended by the government for some people and selling these at a profit to others. To what extent such activities are immoral when they are already illegal is not our concern here. We are concerned instead with the ultimate harm that is thought to follow from allowing forms of "speculation" that are legal in capitalist countries.

Soviet opposition to such speculation might come about as a result of opinions that speculative price increases are unfair, or as a result of opposition to income inequalities that might result from allowing people to speculate, or from the anti-business sentiments that we discussed in the preceding section. However, we have yet to explore a separate issue: whether speculation is viewed as disruptive in that it creates excess price volatility or shortages. Such a view would further justify laws against speculation.

B6. *If the price of coffee on the world market suddenly increased by 30 percent, what do you think is likely to be the blame?* [Response choices: 1) *Interventions of some government*; 2) *Such things as bad harvest in Brazil or unexpected changes in demand*; 3) *Speculators' efforts to raise prices.*]

Response	U.S.S.R.	U.S.A.	$t[3 \text{ vs. } (1+2)]$ (d.f.)
1	17%	13%	-2.93
2	51%	36%	(212)
3	32%	51%	
N:	109	111	

Surprisingly, the Americans were more likely to hold speculators responsible. To put this result into proper perspective, it is worthwhile to note that currently in the Soviet Union the "speculators" are vehemently blamed by the government and certain populist movements for "aggravating shortages" and bringing about price increases. The general public seems to be more skeptical about speculators' capabilities.

This finding was further confirmed by responses to another question that addressed the issue of speculation more directly:

C8. *Grain traders in capitalist countries sometimes hold grain without selling it, putting it in temporary storage in anticipation of higher prices later. Do you think this "speculation" will cause more frequent shortages of flour, bread, and other grain products? Or will it cause such shortages to become rarer?* [Response choices: 1) Shortages more common; 2) Shortages less common; 3) No effect on shortages.]

Response	U.S.S.R.	U.S.A.	$t[1 \text{ vs. } (2+3)]$ (d.f.)
1	45%	66%	1.54
2	31%	26%	(172)
3	24%	8%	
N:	110	112	

Thus, it is true that Soviets tend to blame speculators for shortages, but the Americans do so even more.

Overall, the present survey did not provide evidence that Americans were any more enlightened in their understanding of the functioning of free markets. (For complementary evidence on attitudes toward "profiteering," see question C10 in Section II.)

VIII. Understandings of Compensated Price Changes

At the time this paper was being written (June 1990), there was a heated debate going on in the Soviet Union on about whether

the public would tolerate the compensated increase in the price of bread and other grain products suggested by the Ryzhkov government. While the opinions expressed undoubtedly were heavily motivated by political issues at stake, it was rather disconcerting to hear repeated assertions that a fully compensated price increase was unacceptable because it would adversely affect the standard of living.

Our survey, completed just before the Ryzhkov government put forward its proposal, directly addressed the issue of a compensated price increase:

C6. *Suppose the price of electricity rises four-fold, from 10 cents per kilowatt hour to 40 cents per kilowatt hour. No other prices change. Suppose also that at the same time your monthly income increases by exactly enough to pay for the extra cost of electricity without cutting back on any of your other expenditures. Please evaluate how your overall material well-being has changed. Would you consider your situation: 1) Somewhat better off; 2) Exactly the same; 3) Somewhat worse off?*

Response	U.S.S.R.	U.S.A.	<i>t</i> [1 vs. 3] (d.f.)
1	9%	3%	-2.63
2	77%	63%	(64)
3	14%	34%	
N:	120	121	

Much to our surprise, the responses are consistent with the hypothesis that the Soviets had a better understanding than the Americans that such a change either makes no difference in well-being or improves it.

A related question was asked, outlining a scenario of a compensated increase in the aggregate level of prices:

B10. *Suppose that economists have come to the conclusion that we could substantially improve our standard of living in the next year if we would be willing to accept a 30-per-*

cent inflation rate (increase in the prices of goods by 30 percent). This would mean that our incomes would rise by more than 30 percent. Then we could buy more goods at the new higher prices. Would you support such a proposal?

Response	U.S.S.R.	U.S.A.	<i>t</i> [1 vs. 2] (d.f.)
1) Yes	47%	28%	-3.17
2) No	53%	72%	(226)
N:	118	115	

In accordance with the previous finding, the Soviets proved to be more tolerant of inflation (that was not eroding their incomes) than Americans. (The different answers might also be affected by a difference in the way economists are viewed in the two countries, but the direction of this particular bias is unclear to the present authors.)

IX. Expectations of Possible Future Government Interference

Much recent economic theorizing has emphasized that economic agents respond not only to current government policy but also to anticipated future government policy. Unless the government can commit itself to a new policy, economic agents may, in making long-term decisions, assume that an older policy regime is still relevant. Thus, another impediment to the development of markets in the Soviet Union may be the lingering effect of a memory of the old regime and a feeling that some of its features may be back in the future.

We did find a substantial difference that relates to expectations that the government might usurp the investments people make in private businesses:

C7. *How likely do you think it is that in the next few years the government will, in some way, nationalize (that is, take over) most private businesses with little or no compensation to the owners? Is such nationalization quite likely, possible, unlikely, or impossible?*

[Response choices: 1) Quite likely; 2) Possible; 3) Unlikely; 4) Impossible.]

Response	U.S.S.R.	U.S.A.	$t[(1+2) \text{ vs. } (3+4)]$ (d.f.)
1	20%	5%	-6.37
2	40%	11%	(214)
3	29%	53%	
4	11%	31%	
N: 114		118	

From the Soviet answers here, it would appear that there should be substantial reservations about investing too many resources in cooperatives.

We thought also that Soviets would have a rather weak incentive to save, because of a feeling of insecurity of their savings. After our survey, the Pavlov government actually imposed restrictions on the amount one can withdraw from bank accounts, but our Soviet respondents did not show strong anticipation of such government interference with savings:

B8. *How likely is it, from your point of view, that the government in the next few years will take measures, in one way or another, to prevent those who have saved a great deal from making use of their savings? Is it quite likely, possible, unlikely, or impossible that the government will do this?* [Response choices: 1) Quite likely; 2) Possible; 3) Unlikely; 4) Impossible.]

Response	U.S.S.R.	U.S.A.	$t[(1+2) \text{ vs. } (3+4)]$ (d.f.)
1	17%	15%	-1.34
2	44%	37%	(221)
3	21%	39%	
4	19%	9%	
N: 112		117	

There is some evidence of less confidence of the Soviets, best visible in the "(1+2)/(3+4)" proportion: 61/39 for the Soviet Union; 52/48 for the United States. This difference is not statistically significant, however, and it is well below our prior expectations. Perhaps Americans were thinking of pressures

on the federal government from the deficit and of actions the government might take, such as renegeing on their savings-and-loan obligations, or changing the social-security system or medicare system.

X. Interpretation and Conclusion

It is useful to consider the results of our survey in the context of a specific example of the kinds of things that go wrong in the Soviet Union today. There has been recently a shortage of soap in the Soviet Union. Why has this happened? Why aren't many people setting up cottage industries to manufacture soap (a product that is extremely simple to produce, as industrial commodities go)? Why isn't someone buying soap from available sources and distributing it around the Soviet Union? In short, why aren't the fledgling entrepreneurs in the Soviet Union dealing with the shortage problem?

On one level, the answer is that it is difficult for an enterprise to obtain special permission to start manufacturing or distributing soap. However, on a deeper level, one might ask, why on earth should one need any permission to manufacture and distribute soap in a country that is suffering so much from a shortage of soap? Why should there be any public support for regulators who deny permission for new cooperatives to start to produce or distribute soap?¹⁵

In this paper, we have investigated a number of possible theories to explain why people might feel that the laws should prevent private forces from dealing with the shortage of soap and, hence, why potential private producers of soap might not even try to get the necessary permission or might fear social pressure against such an enterprise. One theory is that people are concerned with fairness of prices, and they would not want to allow prices of soap to

¹⁵Of course, delays to give regulators time to assess the environmental impact of a new manufacturing enterprise may well receive public support, and this particular extreme shortage would likely be viewed as temporary.

rise to reflect the scarcity. Another theory is that people are concerned with the income inequality that might be created if a few entrepreneurs make a lot of money selling soap. Yet another theory is that people do not perceive that the production of soap would be much more effective in a situation where the laws permitted incentives for private production.

While survey questionnaire results do not constitute definitive proof about social attitudes, none of the above-mentioned theories for the relative lack of success of free markets in the Soviet Union has any support in our results. In this study, Soviets appear to be no more concerned with fairness of prices than are U.S. citizens. Further, Soviets appear to be no more concerned with income inequality, and they appear to have the same or even stronger appreciation of the importance of incentives.

Other theories are that there is simply a resistance toward the exchange of money among individuals, as contradicting a sense of regularity in contractual relations, that there is a general lack of interest in starting and running businesses, or that there is a fear that the government will do something in the future to remove the wealth of successful people. We did find some evidence that there is such a resistance toward exchange of money and less warm attitudes toward business; we found also that there may be more of a concern that the government may later nationalize private enterprises. This evidence is of great concern in assessing the long-run outlook for the level of prosperity of the Soviet Union. Still, these differences do not seem so large as to be considered the prime suspects in the annoyingly tangible and immediate problems today, like that of the soap shortage.

Because the differences between the Soviet Union and the United States we found were often small or nonexistent, we feel that perhaps too much prominence has been given in discussions of the transition to a market system in the Soviet Union today to the differences between Soviets and people in market economies. The pressing and immediate problems faced in the Soviet Union

today may be instead political and institutional in nature. When a country inherits an institutional and political framework that has been anti-market, it serves certain entrenched interests in that country to resist change. Thus, individuals who benefit from the present system may make public appeals to fairness, abhorrence of income inequality, and other attitudes to try to stop change. Alternatively, well-meaning Soviet government planners may feel constrained by their incorrect belief that the Soviet public is much more concerned with fairness or income inequality than are the publics in capitalist countries.

Indeed, we have found here that Soviets are concerned with fair prices and with income inequality, so that these concerns might help prevent change to a market economy. However, at the same time, these concerns appear to be little different among Americans. Perhaps Americans would resist perestroika with as much vigor if they inherited the Soviet political and institutional system.

In considering the remarkable similarity between many of the Soviet and American results, it may be well to recall a much earlier interpretation of comparison of Americans with Europeans. Alexis de Tocqueville, in his 1850 book *Democracy in America*, wrote that the "love of money" found among Americans was not a consequence of their national character, but was the natural consequence of a stable system organized around private initiative:

What I say about the Americans applies to almost all men nowadays. Variety is disappearing from the human race; the same ways of behaving, thinking, and feeling are found in every corner of the world. This is not only because nations are more in touch with each other and able to copy each other more closely, but because the men of each country, more and more completely discarding the ideas and feelings peculiar to one caste, profession, or family, are all at the same getting closer to what is essential in man, and that is everywhere the same. In that way they grow alike, even with-

out imitating each other. One could compare them to travelers dispersed through a huge forest, all the tracks in which lead to the same point. If all at the same time notice where the central point is and direct their steps thither, they will unconsciously draw nearer together without either seeking, or seeing, or knowing each other, and in the end will be surprised to find that they have all assembled at the same place. (p. 591)

REFERENCES

- Aage, Hans, "Popular Attitudes and Perestroika," *Soviet Studies*, 1991, forthcoming.
- Abalkin, Leonid, "Too High a Price," *Literaturnaya Gazeta*, 6 June 1990, No. 23 (5297), 9.
- Abercrombie, Nicholas, Hill, Stephen and Turner, Bryan S., *The Dominant Ideology Thesis*, London: Allen & Unwin, 1980.
- Bauer, Raymond A., Inkeles, Alex and Kluckhohn, Clyde, *How the Soviet System Works*, Cambridge, MA: Harvard University Press, 1957.
- Elster, Jon, "Social Norms and Economic Theory," *Journal of Economic Perspectives*, Fall 1989, 3, 99-117.
- Gerschenkron, Alexander, "Social Attitudes, Entrepreneurship, and Economic Development," in Alexander Gerschenkron, ed., *Economic Backwardness in Historical Perspective*, Cambridge, MA: Belknap, 1962, pp. 52-71.
- Kahneman, Daniel, Knetsch, Jack L. and Thaler, Richard, "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review*, September 1986, 76, 728-41.
- Levada, Yu. A., ed., *Yest Mneniye* (I've Got an Opinion), Moscow: Progress Publishers, 1990.
- Lewin, Kurt, Dembo, Tamara, Festinger, Leon and Snedden Sears, Pauline, "The Level of Aspiration," in Joseph McVicker Hunt, ed., *Personality and the Behavioral Disorders*, New York: Ronald, 1944, pp. 333-78.
- McAuley, Alistair, "Social Welfare Under Socialism: A Study of Soviet Attitudes Towards Redistribution," in David Collard, Richard Lecomber, and Martin Slater, eds., *Income Distribution: The Limits to Redistribution*, Bristol, U.K.: Sciencetech-nica, 1980, pp. 238-57.
- Millar, James R., *Politics, Work and Daily Life in the USSR*, New York: Cambridge University Press, 1987.
- Rukavishnikov, V. O., "Ochered'" [The Queue], *Sotsiologicheskkiye Issledovaniya* [Sociological Studies], 1989 (4), 2-12.
- Shiller, Robert J., Boycko, Maxim and Korobov, Vladimir, "Popular Attitudes Towards Free Markets: The Soviet Union and the United States Compared," Cowles Foundation Discussion Paper No. 952, Yale University, August 1990.
- Smith, Tom W., "Inequality and Welfare," in Roger Jowell, ed., *British Social Attitudes: Special International Report*, Aldershot, Hants: Gower, 1989.
- Tocqueville, Alexis de, *Democracy in America*, 1850 Ed., translated by George Lawrence, New York: Harper & Row, 1966.
- Moskva v Tsifrakh, "Finansi i Statistika," Moscow: Statistika, 1989.

Price Distortion and Shortage Deformation, or What Happened to the Soap?

By MARTIN L. WEITZMAN*

The model of this paper generalizes the classical theory of consumer behavior to the more general case of prices that are not necessarily market-clearing. Suppose that, in addition to the money cost, some sort of search, waiting, or other quasi-fixed "effort-cost" is needed to obtain goods. The presence of this quasi-fixed cost element will trigger an inventory policy. A shortage equilibrium occurs when effort costs are such that, in the corresponding inventory policy, the flow of desired consumption does not exceed the available supply flow. Stock hoarding, a critical phenomenon in the economics of shortage, emerges as a natural component of this model. A complete characterization of a stationary shortage equilibrium is given. Comparative statics and welfare analysis are performed. The dynamic transition between steady states is analyzed to give insight into the mechanics of how shortages develop. (JEL D50)

It is known, in a general way, that price distortions lead to shortages, queues, searching, hoarding, and so forth. Yet it seems fair to say that the exact mechanism integrating each main element of a "shortage syndrome," especially the stockpiling phenomenon, has not been clearly articulated.¹ The main aim of this paper is to provide a usable model of shortages by appropriately generalizing the classical theory

of consumer behavior to a situation where prices are not necessarily market-clearing. The model essentially consists of an equilibrium approach that combines inventory theory with demand theory.

A particularly vivid illustration of the phenomenon I have in mind is illustrated by recent Soviet experience in the consumer-goods market. Consider Soviet soap as a metaphorical example. The example is metaphorical because, while the shortage phenomenon in the consumer-goods market I seek to describe is quite general, the particular commodity most illustrative of the general phenomenon can vary.

Throughout most of 1989 (at the time this paper was written), there was virtually no soap available on the shelves of Soviet stores. When officials in charge of planning were asked about this problem, they acted embarrassed and annoyed. In newspaper articles and television talk shows, they explained repeatedly that production this year was actually up 10 percent over last year, which was itself 4 percent higher than the previous year. Furthermore, not only was production accelerated somewhat as this embarrassing shortage became evident, but more than \$8 million of valuable foreign exchange was spent on buying soap abroad. Finally, they pointed out that statistics of

*Department of Economics, Harvard University, Cambridge, MA 02138. This research was supported by a grant from the National Science Foundation.

¹Some models dealing with somewhat different aspects of this phenomenon are described in Janos Kornai and Jorgen Weibull (1978), Victor Polterovich (1983), Dale Stahl and Michael Alexeev (1985), Kent Osband (1989), and the references cited in these works. The case typically treated has the waiting-effort cost proportional to the amount of the good bought. This is equivalent to the assumption that people are limited to buying one small unit at a time and must wait in line anew for each small purchase that is made. (It is then a straightforward exercise to generalize to an "as if" market equilibrium where the "as if" price is the money price plus the appropriately normalized disutility of waiting-effort price.) I much prefer the opposite assumption: after waiting in line for a sufficiently long time, or happening upon the good, the customer can effectively buy as much as he wants. I think this is a more realistic assumption of the two extremes; additionally it leads naturally to an analysis of the key inventory-stock aspect of the problem.

per capita soap consumption show the Soviet Union to be not very far behind the advanced Western capitalist countries. All of this sounds quite believable; and on the whole Soviet citizens do not seem to be going unwashed.

The above story can be repeated for any number of commodities. Soviet leaders frequently characterize the economy as being in a "crisis situation" ("krizisnoye polozheniye"); but what, exactly, is the crisis? What is the appropriate model of causality leading from budget deficits and a monetary overhang to disorders in the consumer-goods markets? What is happening? What should be done?

To such questions no clear answers emerge. Some cite breakdowns in the distribution system. (Railroads seem particularly to be accused.) Others blame a "hoarding psychology" that causes panic buying and is somehow related to the deficit and monetary overhang. Theft by workers, sabotage, and speculation by cooperatives are also candidates. The officials seem unified only on promising increased production to meet the shortage and on calling for formation of committees to investigate formally the problem.

Some informal investigation reveals an interesting, if perhaps not unexpected, fact. Although few official figures are available, observations, conversations, and anecdotes suggest strongly that Soviet people are hoarding soap and other commodities in massive amounts. Significant parts of bathrooms, closets, hallways, and other areas have been given over to storage.

My aim is to model carefully the general process of "hoarding psychology," which is a fairly widespread occurrence in shortage situations, even if it is not typically so extreme as the above case. I believe the model has potential applications to a wide variety of situations where prices are stuck at "wrong" values for whatever reason. Thus, some conclusions may have relevance for malfunctioning markets in capitalist economies and may even help to understand certain features of "fixed-price" macroeconomics. It will be shown that hoarding psychology can be given a quite

rational economic interpretation and can be coherently analyzed within the appropriately extended framework of standard economic theory.

I. The Model

Suppose there are n goods in the economy, denoted $i = 1, 2, \dots, n$. For the sake of argument, all consumers are assumed to be identical. (Allowing consumers to be different would not affect the existence or general form of an equilibrium, but it would render less sharp the characterization of its properties.)

Each consumer has the same utility function of the form

$$(1) \quad U(\mathbf{d}) - V(e) - W(s).$$

In the above formula, $\mathbf{d} = (d_i)$ is the usual consumption n -vector. The variable e_i stands for the amount of "effort" required to obtain good i each time that it is obtained. Conceptually it is perhaps easiest to think of e_i as the time wait in line needed to buy good i . (There is a separate line for each good, and after waiting the required time,² the consumer can buy as much as desired.) However, e_i could more generally represent search effort of any sort expended to obtain the commodity. On the most abstract level, e_i is interpreted as the degree of difficulty in obtaining good i . If good i is obtained at frequency f_i , then the total effort (per unit time) expended on obtaining all goods³ is

$$(2) \quad e \equiv \sum f_i e_i.$$

While it is conceptually easiest to think of effort e_i as deterministic, there is no prob-

²Note that e_i represents the time wait in line, not the length of the line. If m people, each of whom stocks up amount s , are waiting in line and the total flow of goods into the store is d , then the time wait in line is $e = ms/d$.

³By writing disutility as a function of total effort, $V(e)$, I am implicitly assuming that the sum of a large number of fixed costs incurred at different times can, in effect, be smoothed.

lem with an interpretation that makes each e_i an independently distributed random variable so long as it is small relative to n . Then by the law of large numbers, e itself will be (almost) deterministic, equal to a weighted sum of the form (2), where e_i is now interpreted as the *expected* effort that must be spent on obtaining good i .

The variable s_i stands for the stock of good i that is purchased and must be stored when the good is obtained. The coefficient h_i represents the opportunity cost per unit of good i carried (per unit time). The magnitude of h_i would reflect such things as the opportunity cost of storage space i takes up (on shelves, in refrigerators, in warehouses, or whenever applicable), shrinkage, the cost of guarding, interest forgone, the inconvenience of hoarding, and so forth. Total storage cost is then

$$(3) \quad s \equiv \sum h_i s_i.$$

The representative consumer's utility function is assumed to be of the additively separable form (1). The first element $U(\mathbf{d})$ is just the traditional utility function of classical consumer theory, having all the usual properties. The function $V(e)$ represents the disutility of effort, while $W(s)$ is the disutility of storage. The underlying assumption in (1) of independence would appear not to be terribly restrictive, and perhaps even reasonable, in the present context. More general formulations could be treated but with some loss in crispness of results. It is assumed that $U(\mathbf{d})$ is smoothly concave, while $V(e)$ and $W(s)$ are smoothly convex.

The representative consumer faces fixed nominal prices $\mathbf{p} = (p_i)$ and is endowed with nominal money income I . Thus, the usual budget constraint

$$(4) \quad \mathbf{p}\mathbf{d} \leq I$$

holds.

Unlike the classical setup, however, in the situation here prices are not necessarily market-clearing.⁴ Quantities available per

capita are $\mathbf{q} = (q_i)$. Any feasible consumer demand must satisfy the additional constraints

$$(5) \quad \mathbf{d} \leq \mathbf{q}.$$

The supplementary constraints (5) distinguish the present model from classical consumer theory. In the classical case, in effect $q_i = \infty$ for all i , or else \mathbf{p} represents equilibrium prices that just exactly make demands \mathbf{d} equal to supplies \mathbf{q} . For the classical case, in effect there are no explicit constraints on consumer purchases other than the overall budget constraint (4). Here, the interesting case is when constraint (5) "bites" for some goods, representing inadequate supply at the fixed prices, presumably arising ultimately from production limitations.

The present formulation is sufficiently rich to cover a number of special situations of interest. For example, the price might be artificially repressed on only one or a few goods, which then become "deficit" compared to the bulk of commodities, which are market-clearing; or there could be a general deficit of commodities, meaning the prices of most goods are artificially low relative to incomes and availability. Also covered is the case in which the same good is available cheaply in limited amounts at state stores and simultaneously at market-clearing prices in private stores. Yet another situation covered with only slight modification of the present framework is that in which a given fixed vector of goods is to be allocated, so that consumers end up with the same final allocation of goods in any case, but for some given subset of (deficit) goods prices are frozen at below market-clearing levels, while for the remaining (available) goods, prices

⁴It is beyond the scope of this paper to speculate on why certain prices may be set at below market-clearing

levels in certain circumstances. Suffice it to note here that the practice is extremely widespread, and there is an extensive literature on many aspects of it. Governments are fearful of raising prices once they have become established, often with good reason because people do not like price increases. The present paper is limited to analyzing the effects of too low prices without delving deeply into the issue of why they are too low in the first place.

move freely to their competitive levels, which just clear the fixed supplies taking account of consumer income. Since $U(\mathbf{d})$ represents the familiar utility function of goods consumed, it automatically embodies the usual relations of complementarity, substitutability, diminishing returns, and whatever else might be considered relevant to the situation at hand.

II. Coefficient of Price Distortion

In what follows, it will be useful to have a quantitative measure of the degree to which values and prices are distorted in the economy under consideration. To that end, consider the following mathematical programming problem:

$$(6) \quad \max[U(\mathbf{d})]$$

subject to

$$(7) \quad \mathbf{d} \leq \mathbf{q}$$

$$(8) \quad \mathbf{p}\mathbf{d} \leq I.$$

The constrained optimization (6)–(8) is a classical resource-allocation problem. In the present context, it can be interpreted as a second-best problem in optimal rationing. Let the solution be

$$(9) \quad \mathbf{d} = \mathbf{d}^*.$$

Let λ be the shadow price of constraint (8). Without significant loss of generality, suppose that the marginal utility of income is positive or that $\lambda > 0$. [The marginal utility of an extra ruble is greater than zero, which means that constraint (7) is not so tight in every component that no goods are available to be bought on the margin. This would be guaranteed in theory if, for example, some of the goods were available in unlimited supply, or if prices on some of the goods were market clearing.] Then, it is not difficult to see that

$$(10) \quad \lambda = \min_i \left[\frac{U_i}{p_i} \right]$$

where

$$(11) \quad U_i \equiv \left. \frac{\partial U}{\partial d_i} \right|_{\mathbf{d} = \mathbf{d}^*}.$$

Necessary and sufficient conditions for an optimum are then

$$(12) \quad U_i > \lambda p_i \rightarrow d_i^* = q_i.$$

While λp_i represents the nominal price of an extra unit of good i (normalized so that the marginal utility of income is 1), U_i measures the actual value of an additional unit of good i , or what people would actually be willing to pay. In the present context, it is then natural to define the coefficient of value distortion or price distortion of good i (weighted by the amount of the good) as

$$(13) \quad \delta_i \equiv (U_i - \lambda p_i) d_i^*.$$

The coefficient δ_i measures the difference between the actual value of good i that might be consumed and the nominal price value, normalized in terms of utility. Therefore, δ_i is a measure of the degree of disequilibrium deviation from market-clearing of good i . If the logic of (13) is accepted, the appropriate measure of overall value or price distortion in the economy becomes

$$(14) \quad \delta \equiv \sum \delta_i.$$

The coefficient δ is a measure of the degree of overall disequilibrium in the economy.

III. The Basic Problem

The question now arises as to how the goods are actually distributed in the model economy, given the existence of constraints (4) and (5). As a point of departure, suppose that a state of chronic shortages exists in an orderly stationary equilibrium. That is, every consumer knows he must expend effort e_i to obtain good i and plans the appropriate inventory policy. After, in effect, paying the fixed waiting-time cost of e_i , the consumer chooses to buy and stock the amount s_i and run it down at the consumption flow rate d_i . This pattern is repeated at

frequency

$$(15) \quad f_i = \frac{d_i}{s_i}.$$

Furthermore, economy-wide this repetitive behavior is self-reinforcing.

Of course this description of shortage behavior as a regular steady-state equilibrium with recurrent sawtooth-patterned inventories is an abstraction. Shortage phenomena can be notoriously erratic. Nevertheless, treating a shortage economy as if it were in a stationary equilibrium yields important quantitative insights. Furthermore, it is a necessary first step to any analysis of dynamics. Actually, the methodological issues connected with modeling a shortage equilibrium do not seem fundamentally different from those involved in modeling a nonshortage equilibrium.

Consider the problem facing the typical consumer. Effort levels $\{e_i\} \geq 0$ are taken as given.⁵ Consumption flow levels $\{d_i\}$ and inventory stocks $\{s_i\}$ should be chosen to:

$$(16) \quad \underset{\{d_i, s_i\} \geq 0}{\text{maximize}} \left[U(\{d_i\}) - V \left(\sum e_i \left[\frac{d_i}{s_i} \right] \right) - W(\sum h_i s_i) \right]$$

subject to

$$(17) \quad \sum p_i d_i \leq I.$$

In what follows, I assume that the first-order necessary conditions for characterizing an optimum to the above problem [(16) and (17)] are also sufficient.⁶

⁵In equilibrium, $\{e_i\}$ will be like an implicit price that equilibrates the system, but each individual consumer will view $\{e_i\}$ as exogenously given.

⁶A variety of conditions would guarantee this result. Essentially, the inventory part is introducing an economy of scale into an otherwise convex problem. So long as the inventory nonconvexity effect can be bounded [e.g., by assuming that $\{e_i\}$ is small relative to the curvature of $U(\cdot)$], the necessary first-order conditions for (16)–(17) would remain sufficient. In this paper, I will not go further into the essentially technical and messy aspect of insuring that necessary conditions are

In the economy being modeled, the appropriate equilibrium concept is the following.

DEFINITION: A stationary shortage equilibrium is a set of $\{e_i, d_i, s_i\}$ satisfying the following three conditions:

$$(18) \quad \{d_i, s_i\} \text{ solves problem}$$

$$(16) \text{--}(17) \text{ for given } \{e_i\}$$

$$(19) \quad d_i \leq q_i \text{ for all } i$$

$$(20) \quad \text{if } d_i < q_i, \text{ then } e_i = 0.$$

The reader should feel satisfied, upon reflection, that conditions (18)–(20) represent the correct generalization of classical consumer equilibrium theory to the present context.⁷

It is not difficult to generalize the definition of a stationary shortage equilibrium to a situation with many nonidentical consumers having different utility functions and different incomes, nor is it difficult to prove existence using methods similar to those employed in the present paper. This route is not pursued in detail here simply because, as with most truly general equilibrium formulations, it is impossible to characterize sharply the properties of a solution without placing more structure on the model.

The following theorem completely characterizes a stationary shortage equilibrium.

sufficient for the given problem. I am indebted to Victor M. Polterovich for pointing out to me that some assumption needs to be made in order to presume that the necessary first-order conditions for a maximum of (16)–(17) are also sufficient in the present context.

⁷There is an implicit assumption behind definition (18)–(20) that at any instant in time consumers hold uniformly distributed stocks of good i , ranging from s_i to 0, and they arrive uniformly to market just as their stock of the good is depleted. This assumption could presumably be justified as the outcome of a dynamic optimizing process in which consumers satisfying the overtaking criterion always choose the line with the shortest wait and thereby force all waiting lines for the same good to have equal length in equilibrium.

THEOREM 1: *The unique stationary shortage equilibrium is the solution of the following equations.⁸*

$$(21) \quad d_i = d_i^*$$

$$(22) \quad s_i = \frac{\delta_i}{wh_i}$$

$$(23) \quad e_i = \frac{\delta_i^2}{vwh_id_i^*}$$

where

$$(24) \quad w \equiv W'(s)$$

is the marginal disutility of storage evaluated at $s (= \sum h_i s_i)$ satisfying

$$(25) \quad sW'(s) = \delta$$

while

$$(26) \quad v \equiv V'(e)$$

is the marginal disutility of effort evaluated at $e (= \sum f_i e_i)$ satisfying

$$(27) \quad eV'(e) = \delta.$$

PROOF:

In the optimization problem (16)–(17) that defines condition (18), let $\mu \geq 0$ be the shadow price multiplier for inequality (17). The corresponding necessary and, by assumption, sufficient first-order conditions for any $d, s \geq 0$ satisfying (17) to be the unique maximizer of (16) are, for all i ,

$$(28) \quad U_i - V'\left(\frac{e_i}{s_i}\right) = \mu p_i$$

$$(29) \quad V'\left(\frac{e_i d_i}{s_i^2}\right) = W'(h_i).$$

⁸ Conditions (21)–(27) are presented in the given sequence and form to facilitate their economic interpretation. From the strictly mathematical standpoint of presenting an algorithm that uniquely solves (18)–(20), it is preferable to think of the following order: first (21) defines $\{d_i\}$, then (25) defines s and (27) defines e , then (24) defines w and (26) defines v , then (22) defines $\{s_i\}$ and (23) defines $\{e_i\}$. Equations (25) and (27) have unique solutions in s and e , from the convexity of $V(e)$ and $W(s)$.

[Equation (29), rearranged, is the famous square-root law of inventory theory.]

The rest of the proof is essentially by inspection. Assignments (21)–(27) and the supplementary assignment

$$(30) \quad \mu = \lambda$$

are proposed as the solution of (18)–(20). Using conditions (4), (5), (12), and (13), it is straightforward to verify that the proposed solution, (21)–(27) and (30), does indeed satisfy (28), (29), (19), and (20).

With the shortage equilibrium expressed in the simple closed form (21)–(27), it is easy to perform comparative-static exercises.

Note from (22) that s_i is proportional to δ_i , while from (23) e_i is proportional to δ_i^2 . Thus, small shortages show themselves primarily in increased stock hoarding, with just very small increases in search activity. On the other hand, large shortages result in large inventories and very large waiting lines or search times. These observations suggest how shortages might evolve or devolve.

While from (23) effort per purchase is proportional to the square of the coefficient of price distortion, total effort is not. This is because, as waiting lines increase, the consumer reacts by buying bigger bundles less frequently. The total effort per unit time spent on obtaining good i is, from (15) and (21)–(23),

$$(31) \quad f_i e_i = \delta_i / v.$$

Substituting (31) and (22) into (13) yields, respectively,

$$(32) \quad U_i = \lambda p_i + v \frac{f_i e_i}{d_i}$$

$$(33) \quad U_i = \lambda p_i + w \frac{h_i s_i}{d_i}.$$

Thus, the difference between the intrinsic value of a good and its nominal price is made up by the effort expended per unit of the good to obtain it [equation (32)] and also by the cost expended per unit of the good to store it [equation (33)]. In shortage

equilibrium, the consumer ultimately gets what goods [equation (21)] he would have gotten under the optimal rationing scheme (6)–(8). However, from (32) and (33), it is the cost of nonproductive search and storage activity that raises the “as if” equilibrium price from λp_i to U_i , the hypothetical price that would have to be paid to clear the “as if” competitive market.

In the simplified world of this model, it is easy to analyze the welfare loss of a distorted price system. As a reference point, I will take the constrained second-best optimal-rationing solution (9) of problem (6)–(8). Such an allocation maximizes utility subject to budget constraint (4) and goods-availability constraint (5). In some sense, the utility difference between it and the shortage equilibrium, to be denoted L , represents the waste of search and storage activity. There is a simple expression linking L with the coefficient of price distortion δ .

THEOREM 2: *The welfare loss of a stationary shortage equilibrium compared with an optimal rationing scheme is*

$$(34) \quad L = \delta \left(\frac{1}{a} + \frac{1}{b} \right)$$

where a is the elasticity of disutility of effort

$$(35) \quad a \equiv \frac{v}{V} e$$

while b is the elasticity of disutility of storage

$$(36) \quad b \equiv \frac{w}{W} s.$$

PROOF:

From (21), the allocation of goods is the same in (18)–(20) as in (6)–(8). Therefore, from (1) the difference in utility attained is

$$(37) \quad L = V(e) + W(s).$$

Making use of definitions (25), (27), (35), and (36), expression (37) can be rewritten as (34).

Note that shortage losses are first-order in the appropriately normalized measure of

economy-wide price distortion. This is because, unlike the standard second-order deadweight loss (from tax theory) of distorted prices in markets that clear, here malformed prices cause real shortage deformations akin to rent-seeking activity.⁹ Furthermore, the social cost of these shortage deformations enters twice; that is, distorted prices here do double damage. First, they create nonproductive search activity at a social cost of δ/a , which has no other function than to allocate the artificially underpriced goods. Second, price distortions result in goods being tied up by buyers at social cost δ/b in socially unnecessary inventories held throughout the system.¹⁰

Somewhat paradoxically, the higher the elasticity of disutility of effort or storage, the less damage is done by a faulty price system. This is because search or storage activities do not play a directly productive role here. Their only purpose is to allocate artificially underpriced goods. Since buyers can only end up consuming on average what is being made available on average, the goods will end up getting distributed the same way no matter what are the $V(e)$ or $W(s)$ functions. When people suffer greater incremental pain from additional search or storage activity, they will actually end up with less total disutility, because they will simply avoid waiting in lines or stocking up their closets. Hence, the paradoxical conclusion is that greater potential pain is less actual pain.

⁹For an analysis of rent-seeking activity, see James Buchanan et al. (1980) and the references cited therein.

¹⁰The reader may wonder why distorted prices do double damage in this model. For concreteness, take the linear case $a = b = 1$. Then, (34) becomes $L = 2\delta$. Yet in a standard model where the waiting-effort cost is proportional to the amount of the good bought, the total loss of consumer surplus would be equal to the price subsidy times the quantity purchased, or δ . Where, then, does the extra term of δ come from? In a standard model, there is no need to hold inventories. In the present square-root inventory model, the marginal cost of carrying inventory is always equal to half the average cost. Since equilibrium essentially requires that the marginal cost of holding inventory be equal to the amount of price subsidy, it follows that the average cost of holding inventory must equal twice the price subsidy. I am indebted to Paul R. Milgrom for providing this intuitive explanation of why $L = 2\delta$ when $a = b = 1$.

There is another seeming paradox here, which makes an important point. In an environment of widespread shortage due to a malfunctioning price system, ordinary notions of "real income" lose their meaning. Actually, if there is sufficient price distortion, higher income can, other things being equal, mean lower welfare. If enough goods are in shortage, more "real income" can translate primarily into longer lines and greater hoarding rather than increased consumption per se.

Other things (including prices) being held equal, when nominal income is raised there are two effects. More money is available to spend on nondeficit goods ($i|U_i = \lambda p_i$), which increases welfare. This is the usual sense in which higher real income is better. However, in a shortage situation there is also a detrimental side, which might be called the "money-overhang effect." As income is raised, the marginal utility of income (λ) declines, which increases the difference between value and price ($U_i - \lambda p_i$) for deficit commodities, which in turn leads to greater search (e_i) and storage (s_i) effort and raises the welfare loss L . This money-overhang effect is more pronounced as the deficit commodities constitute a greater fraction of all goods. In the extreme case, the money-overhang effect can be so severe that more income can actually result in lower welfare.

There is a slightly different way of constructing and interpreting the model that makes the same basic point. Take as given a fixed vector of goods q to be allocated. Consumers end up with the same final allocation of goods $d = q$ no matter what. For a given subset of (deficit) goods, prices are frozen at below market-clearing levels. For the remaining (available) goods, prices move freely to their competitive levels, which just clear the fixed supplies, taking account of consumer income. As money income increases in this setup, the consumers will be made unambiguously worse off. Price distortion on the available goods will always be zero. However, for the deficit commodities, the coefficient of price distortion will increase with income as the marginal utility of income declines. For the same final allocation

of goods, lower fixed prices or higher nominal income decreases consumer welfare.

Thus, standard measures of "real income" may be a quite unreliable indicator of welfare in situations of economic shortage. It is then not difficult to understand the temptation to eliminate search and storage costs by imposing rationing, even though such measures introduce problems of their own.¹¹

In the model as presented, the production or supply side is taken as more or less given. If the supply side were adversely affected by shortages, either because people are so busy finding and keeping goods that they have less time to work or because it is difficult to maintain morale on the job when pay buys so little, the ingredients are present for a vicious circle.¹² An increased money overhang decreases welfare, which then feeds back to lower production, which depresses welfare further. In such a world, income-increasing tendencies that exacerbate a money overhang can be very dangerous indeed.

This point can be illustrated rather simply as follows. Suppose that a unit of labor-time input produces a unit of homogeneous output. Each person is endowed with one fixed unit of labor-time, which can be divided continuously between producing output or engaging in non-directly-productive search and storage activity. The delivery price of the single good (or, in an alternative interpretation, the wage) is fixed by the government at ω , whereas the sales price of the good is fixed by the government at $p < \omega$. Thus, the subsidy of cost over price per unit of the good is $\gamma \equiv \omega - p$. Suppose each person must pay a fixed per capita tax of $\theta > 0$ to the government. Let y be the amount of the good produced per capita. (The remaining time $1 - y$ is spent on search and storage activity.) Then per capita dis-

¹¹ Throughout this paper, black-market activities are ignored. It could be argued that pressure toward black markets might increase with the degree of price distortion.

¹² Indeed, the Soviet economy is showing signs of this vicious circle.

possible income is $\omega y - \theta$, which in equilibrium must equal the value of goods purchased py . Rewriting this latter relation gives $y = \theta/\gamma$. Other things being equal, the greater the price subsidy γ , the greater is the incentive to engage in nonproductive search and storage activity, and the lower is per capita production (and consumption) of actual goods and services. Note too that other things being equal, a higher per capita tax (within the relevant range) actually raises productive output, because it lessens the monetary overhang caused by price subsidization. It should be readily apparent that if this kind of economy gets caught in any kind of spiral between lower standards of living and increased money-wage subsidization, production can implode.

For the sake of a sharp characterization, the analysis has proceeded as if all consumers are identical. A shortage equilibrium will typically exist in the more general case when consumers are different, but it will be more complicated to analyze. Consumers with less income or a lower value of time will end up waiting in line and storing the subsidized products, while people with higher incomes or a greater value of time will tend to seek out the higher-priced goods with shorter lines.¹³

The model of shortage equilibrium presented here presupposes a regular, stationary, well-behaved process. Actually, shortage phenomena frequently display an erratic, nonstationary aspect. To that extent, the analysis presented here probably understates the social loss of price distortions.

IV. Dynamics

The analysis so far has been purely static. A stationary shortage equilibrium is postulated. Everyone expects this steady state to persist forever. Consumer behavior, which is contingent upon this stationary equilib-

rium, reinforces it. As has been shown, the analysis of steady-state shortage equilibrium is eminently tractable. Some suggestive insights emerge about changes in behavior across steady states, but there has been no dynamic analysis as such.

In a model such as this, where inventories are playing a central role, dynamic transitions display an important property not visible from examining steady-state behavior alone. With a change between steady states comes a change in average inventory stocks, which in turn forces a corresponding alteration of consumption flows during the interim. Adjustments toward higher stocks of equilibrium inventories, without a change in supply, must involve an accumulation at the expense of consumption, a kind of "inventory squeeze" of consumption. This in turn leads to an overshooting property in the waiting time of queues, so that short-run welfare temporarily falls below its long-run equilibrium level. The purpose of the present section is to explore this theme in some detail. It turns out that a formal analysis of dynamics for the general case is an extremely formidable task. All that I am able to do here is to provide a suggestive example of the dynamics of one simple case. I believe this example captures well the main elements that are involved in the dynamic properties of an "inventory squeeze." I also believe that the basic features of the example must generalize. However, disequilibrium dynamics is a notoriously difficult area about which rather little is known except that a lot of different things typically can happen. I do not want to claim that all adjustment processes will be as smooth as the example I will describe in detail.

Suppose, then, that the economy starts out in some steady-state shortage equilibrium. This old steady-state equilibrium had been expected to continue indefinitely. Then, suddenly and without warning, some underlying parameters change, which increase the degree of shortage in the economy. To be specific, suppose δ_i increases to δ'_i , but the underlying supply flow of goods does not change. The easiest thing to envisage is an increase of nominal income that exacerbates the degree of monetary over-

¹³ It is even theoretically possible that low-income, low-value-of-time people could be made better off by price subsidies, although this possibility is excluded in a representative consumer world.

hang in the economy. Suppose the new situation is naively expected to last forever.

Using (21)–(23), it is easy to analyze the differences between the two steady states. With $\delta'_i > \delta_i$, the new stationary equilibrium will feature more search effort and larger inventories. Yet this cannot be the entire story of the transition. When the underlying production structure has not changed, consumption will be the same (equal to production) in both steady states, or $d'_i = d_i$. If consumption in the new stationary equilibrium is to be the same as the old, because what has changed is not the underlying production structure but the price of good i relative to income, the economy must go through a wrenching dislocation in making a transition between steady states. There is no way that inventory stocks can be built up from an average per capita level of $s_i/2$ to a higher average per capita level of $s'_i/2$ and have consumption flows remain uninterrupted throughout. With consumption levels the same in beginning and final steady states, but inventory levels higher in the final state, there must be a transition period during which consumption is curtailed to allow stocks to accumulate. In turn, the only allocation mechanism available to force lower consumption is increased waiting time or search costs.

A change from a lower to a higher state of equilibrium shortage involves a change from a lower to a higher level of search time. The greater required effort of the new steady state is bad news enough. The effort level during the transition period is even more awful news because, to induce consumers to curtail consumption and permit inventory stocks to accumulate, effort must “overshoot” its new long-run equilibrium and then only gradually decline to the new steady-state level, which is merely worse than the old.

Suppose the change occurred at time $t = 0$. At time $t < 0$, everyone was expecting the old shortage equilibrium to last forever. Then, at time $t = 0$, δ_i unexpectedly increased to δ'_i . This new condition, too, is expected to be permanent. The typical consumer, when he now goes to wait in line to

make his purchases and restock his inventories at time $t \geq 0$, will attempt to adjust. Projecting the same effort level to obtain the good but at a higher value relative to price, the consumer will initially attempt to buy and to consume more of the good. Since the flow of desired purchases would then exceed the flow of available goods and since price is fixed (by assumption), the length of the queue must increase to throttle back desired purchases to available supply. Even should waiting lines have increased enough to trigger the new steady-state inventory policy, in which desired consumption flow is equal to available supply flow, that is not enough. In the new equilibrium, each consumer wants to hold more inventories and make purchases less often because the ratio of search effort to nominal price is higher than before. That means that every buyer coming to market in the time immediately after $t = 0$ wants to buy more than before in order to stock up to the new inventory level. Then, waiting times must be even higher than the new equilibrium level, to keep potential purchasers away. The effort level to obtain the good in the transition period must be so high that potential buyers are forced to delay purchases by cutting back on consumption and waiting it out until the lines come down to the new equilibrium level.

In Figure 1 are depicted typical “before,” “during,” and “after” patterns of effort, inventory, and consumption. At least this is the pattern of the example I wish to present. Before the shock at $t = 0$, inventory stocks for a representative consumer display the usual sawtooth pattern with periodicity

$$(38) \quad T \equiv \frac{s}{d}.$$

(For notational convenience, the subscript i will henceforth be dropped whenever its use is superfluous from the context.) After adjustment, the new steady-state inventory sawtooth has periodicity

$$(39) \quad T' \equiv \frac{s'}{d}.$$

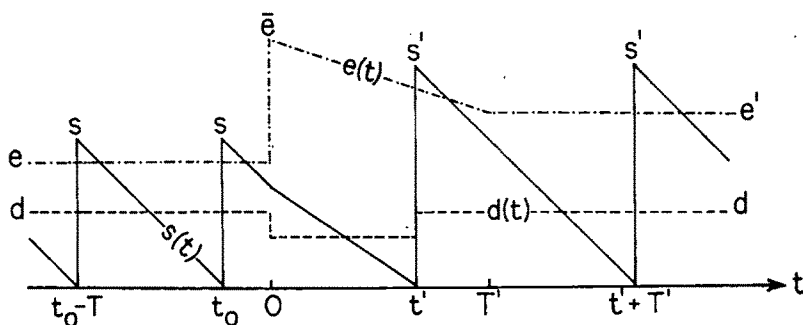


FIGURE 1. TYPICAL TRAJECTORIES OF EFFORT, CONSUMPTION, AND INVENTORY

I will exhibit one relatively simple dynamic transition path that lasts exactly T' time units. At time $t < 0$, the economy is in the old steady-state equilibrium. During time $0 \leq t \leq T'$, the economy is in a state of transition. For times $t > T'$, the economy is in its new steady-state equilibrium. The transition itself is a dynamic equilibrium because everyone believes it will happen this way, and their subsequent optimizing behavior makes it actually come about. There may be other consistent transition paths, but I believe they must demonstrate the same basic "inventory squeeze" aspect of the path I will describe, only in more complicated form, because average inventory stocks cannot be built up without somehow squeezing consumption flows in the process. I am not bothered by limiting myself to the relatively simple example of a transition described here because: a) it is already complicated enough, b) it illustrates well the basic principles involved, and c) there is little purpose to misplaced generality here when the model itself and the other aspects of dynamic behavior are purposely being simplified as much as possible.

At time 0, it comes as a complete surprise when δ changes to δ' . Suppose that after $t = 0$ everyone believes that the effort required to obtain good i has jumped up discontinuously to level \bar{e} ($> e'$) and is declining linearly back down to level e' during the transition period $[0, T']$, thereafter to remain forever at the new steady-state level e' . This effort profile is depicted in Figure 1.

Also shown is a typical consumer's corresponding inventory stock and consumption flow trajectories.

When the shock hits at $t = 0$, the consumer is unexpectedly caught somewhere in the previously optimal inventory cycle. For all $t < 0$, inventories had been picked up and stored in batches of size s , to be run down at consumption rate d . Suppose at $t = 0$ a consumer has on hand stock σ ($0 \leq \sigma \leq s$). In other words, this consumer made his last purchase at time

$$(40) \quad t_0 = \frac{s}{d} \left(\frac{\sigma}{s} - 1 \right).$$

From $t = 0$ on, the consumer faces the declining effort profile depicted in Figure 1. There is then an incentive to delay the next purchase time by slowing down current consumption, because the further off a purchase can be put into the future, the lower will be the effort cost of the purchase.

In the proposed adjustment, every consumer is induced to lower consumption from rate d to rate

$$(41) \quad d' \equiv d \left(\frac{s}{s'} \right)$$

until stocks run down to zero at time

$$(42) \quad t' \equiv \frac{\sigma}{d'} = \frac{\sigma s'}{ds}.$$

From time t' on to infinity, this typical con-

sumer repeats the new steady-state cycle: inventories of size s' are picked up and stored with periodicity T' [equation (39)], being run down by consuming at rate d .

Thus, the typical consumer with inventory σ at $t = 0$ suffers a decline in consumption by the fraction s/s' over his adjustment period $[0, t']$. By time $t = T'$, all consumers are synchronized in the new steady-state.

To prove that the proposed time profiles $e(t)$, $s(t)$, and $d(t)$ depicted in Figure 1 represent a consistent dynamic transition path, two conditions must be shown to hold. First, the feasibility of these inventory and consumption patterns must be demonstrated. Then it must be proved that each consumer wants to follow this prescribed pattern.

Feasibility means that at any instant the flow of supply is equal to the amount of stock being demanded per unit time. The proposed trajectory is feasible for $t < 0$ and for $t > T'$ because these are, respectively, old and new steady-state policies satisfying (18)–(20). For $t \in [0, T']$, a “spreading out” effect occurs. Buyers arriving to market now are each carrying away larger stocks than before (s' instead of s), but these same buyers are arriving less frequently because they are running down existing stock at a rate of d' , which is lower than d . It turns out that the two effects exactly cancel each other, and in aggregate buyers are taking away the same amount per unit time as in the steady state.

In the old steady state at $t < 0$, a consumer with stock σ at time 0 would have planned on next picking up stock s at time

$$(43) \quad \tau \equiv \frac{\sigma}{d}.$$

After the shock, a consumer with stock σ at time 0 will next pick up stock s' at time t' defined by (42). The ratio of the frequency of buyer arrivals in the old equilibrium to the frequency of buyer arrivals in the transition interval is thus

$$(44) \quad \frac{t'}{\tau} = \frac{s'}{s}.$$

Equation (44) is exactly the ratio of the stock each buyer purchases during the transition interval to the stock each buyer purchases in the old equilibrium. Thus, the total stock demand per unit time is the same, equal to supply, and the proposed trajectory is feasible.

Next it must be shown that, given the effort profile $e(t)$ depicted in Figure 1, each consumer desires to follow the prescribed consumption and stock trajectories $d(t)$ and $s(t)$. For the steady states of $t < 0$ or $t > T'$ this has already been shown by (21)–(23). To show that the effort profile $e(t)$ of Figure 1 induces the depicted pattern of $d(t)$ and $s(t)$ for $t \in [0, T']$, I must make some assumptions about consumption payments and inventory costs in the transition period. I do not think that these assumptions are critical in the sense that they are being made primarily to serve as sufficient conditions for the relatively neat patterns of Figure 1 to emerge. I believe that more general assumptions of a reasonable sort would give the same qualitative features but would be harder to analyze.

In what follows, the consumer's basic decision variable is x , the rate of consumption from $t = 0$ until the inventory stock σ runs out, at which time it is optimal to follow the new steady-state policy. Assume that the money cost of consuming any good is pay-as-you-go. It is proposed that consumption is at rate x for $0 \leq t \leq \sigma/x$ and at rate d for $t > \sigma/x$. The utility cost of following this policy over the interval $[0, T']$ is

$$(45) \quad \lambda p x \left(\frac{\sigma}{x} \right) + \lambda p d \left(T' - \frac{\sigma}{x} \right).$$

Assume further that inventory storage cost is essentially the storage-cost coefficient times the new inventory level times the interval over which it operates. Some stories could be told to support this interpretation, but it is more important to emphasize that virtually any reasonable formulation would yield qualitatively similar results. The inventory-carrying cost of the proposed policy over the interval $[0, T']$ is

then

$$(46) \quad hs' \left(T' - \frac{\sigma}{x} \right).$$

For notational convenience, assume that w and v are normalized to unity.

With $e(t)$ representing effort as a function of time, the total net utility of the proposed policy over the interval $[0, T']$ is then

$$(47) \quad \left(\frac{\sigma}{x} \right) U(x) + \left(T' - \frac{\sigma}{x} \right) U(d) - e \left(\frac{\sigma}{x} \right) \\ - \left[\lambda p x \left(\frac{\sigma}{x} \right) + \lambda p d \left(T' - \frac{\sigma}{x} \right) \right] \\ - hs' \left(T' - \frac{\sigma}{x} \right).$$

The consumer chooses x to maximize (47), yielding the first-order condition

$$(48) \quad \frac{\sigma}{x} U_i(x) - \frac{\sigma}{x^2} U(x) + \frac{\sigma U(d)}{x^2} \\ + e' \left(\frac{\sigma}{x} \right) \frac{\sigma}{x^2} - \frac{\lambda p d \sigma}{x^2} - \frac{hs' \sigma}{x^2} = 0.$$

The expression $e'(\sigma/x)$ stands for the time derivative of the effort function evaluated at $t = \sigma/x$.

For the consumer to be choosing to follow the prescribed path of Figure 1, condition (48) must hold for the value

$$(49) \quad x = d'$$

where d' is defined by (41). Substituting (49) and (41) into (48) and rearranging yields

$$(50) \quad e' \left(\frac{\sigma s'}{ds} \right) = U \left(\frac{ds}{s'} \right) - U(d) \\ - \frac{ds}{s'} U_i \left(\frac{ds}{s'} \right) + \lambda p d + hs'.$$

As the parameter σ varies from 0 to s , condition (50) defines the slope of the required effort function $e'(t)$ on the interval

$[0, T']$. Since the right-hand side of (50) does not contain σ , the slope of the effort function, hereafter denoted e' , must be constant throughout $[0, T']$. In other words, $e(t)$ is a straight line, as depicted in Figure 1.

From applying (32) to (50),

$$(51) \quad e' = 0 \quad \text{for } s' = s.$$

In other words, a transition from an old steady state to a new steady state that is the same as the old steady state is no transition at all.

The concave function

$$(52) \quad U(y) - \lambda p y - \frac{hs'y}{d}$$

has a unique maximum, where its derivative is zero, at $y = d$ defined by the first-order condition (32). In particular, this means that, for $y = ds/s'$,

$$(53) \quad U(d) - \lambda p d - \frac{hs'd}{d} \\ \geq U \left(\frac{ds}{s'} \right) - \lambda p \left(\frac{ds}{s'} \right) - \frac{hs'}{d} \left(\frac{ds}{s'} \right).$$

Substituting (53) into (50) and rearranging yields

$$(54) \quad e' \leq - \frac{ds}{s'} \left[U_i \left(\frac{ds}{s'} \right) - \lambda p - \frac{hs'}{d} \right].$$

From concavity of expression (52), the derivative with respect to y of (52) is greater than zero for $y < d$. Letting $y = ds/s'$, this observation applied to (54) yields

$$(55) \quad e' < 0 \quad \text{for } s' > s.$$

Since e' is continuously differentiable in s' , (51) and (55) imply that, when s' is sufficiently close to s ,

$$(56) \quad e' > 0 \quad \text{for } s' < s.$$

It has been shown, then, that in undergoing a change from an old steady state with a lower degree of price distortion to a new

steady state of higher price distortion, there will occur a transition phase in which effort is greater and consumption is lower than in either steady state. Conversely, during a transition to a state of less price distortion, effort will be lower, and consumption will be higher.

All of this means that increased price distortion has particularly painful immediate effects as goods are squeezed out of consumption and into inventory stocks. The situation may stabilize later of its own accord to a still unsatisfactory level, but the period immediately after an increase in monetary overhang is likely to be one of especially acute shortage symptoms.

The good news is that if price distortion can be lessened then society can go on a temporary binge as unwanted inventories are worked off into extra consumption. Reversing the shortage process—by raising prices or lowering money incomes—results in a “consumption dividend” from unneeded stocks coming out of storage into general circulation.

V. Conclusion

This paper has presented a formal model of consumer behavior under conditions of shortage. A critical role in allocating shortage goods is played by the directly unproductive activities of search and storage. It is possible to express directly these two key variables as simple functions of the underlying degree of price distortion and, therefore, to analyze easily the relation between price distortion, search effort, and goods stockpiling. “Hoarding psychology,” or what might better be called “defensive hoarding,” can be thoroughly analyzed as an economic phenomenon by extensions of standard economic theory. In shortage equilibrium, everyone must expend effort to locate and hoard goods because everyone else is expending effort to locate and hoard goods.

A clear theme of the paper is that price distortion and monetary overhang can present very severe threats to the normal functioning of an economy. The essential inadequacy is in the monetary domain of prices and income, not in the real economy of production and distribution. The seeming paradox of the missing Soviet soap must be resolved not by scapegoating distributors or by making token increases in soap production, which do little to alleviate the underlying problem. Instead, prices on soap must be increased, or incomes lowered, so that consumers can move toward a better state where they are not impelled to hoard large inventories. The essential issue is to remove the incentives that lead to excessive inventory stocks blocking what should be a direct flow of goods from production to consumption.

REFERENCES

- Buchanan, James M., Tollison, Robert D. and Tullock, Gordon, eds., *Toward a Theory of the Rent-Seeking Society*, College Station, TX: Texas A&M Press, 1980.
- Kornai, Janos and Weibull, Jorgen W., “The Normal State of the Market in a Shortage Economy: A Queue Model,” *Scandinavian Journal of Economics*, March 1978, 80, 375–98.
- Osband, Kent, “Economic Crisis in a Reforming Socialist Economy,” unpublished manuscript, November 1989.
- Polterovich, Victor M., “Problema Izmereniya Defitsitnosti Blag” [The Problem of Measuring Welfare Under Shortage], *Ekonomika i Matematicheskie Metodi*, May 1983, 19 (5), 878–91.
- Stahl, Dale O. and Alexeev, Michael, “The Influence of Black Markets on a Queue-Rationed Centrally Planned Economy,” *Journal of Economic Theory*, April 1985, 35, 234–50.

A Natural Experiment in the Organization of Entry-Level Labor Markets: Regional Markets for New Physicians and Surgeons in the United Kingdom

By ALVIN E. ROTH*

The histories of seven regional markets for new physicians and surgeons in the United Kingdom are considered. Like the American market, these markets have experienced failures that led to the adoption of centralized market mechanisms. Because different regions employ different centralized mechanisms, these markets provide a test of the hypothesis that the success of the American market is related to the fact that it produces matches which are stable in the sense that no two agents mutually prefer to be matched to one another than to their assigned partners. Even in the more complex U.K. markets, this kind of stability plays an important role. Centralized markets that produced unstable matches in environments in which agents could act upon instabilities fared no better than the decentralized markets they replaced. (JEL C78, D00, J41, J44)

In this paper, I seek to analyze a unique natural experiment that has emerged over the last 20 years in the United Kingdom, concerning the organization of entry-level labor markets. The markets in question are those in which newly graduated medical students seek their first hospital positions, called preregistration house-officer positions. These positions are closely comparable to first-year intern positions in American hospitals (although there are some important differences, which will be

discussed). The natural experiment arises because these markets are organized differently in different regions of the National Health Service. These different markets allow investigation of how market behavior is influenced by market organization. This diverse set of markets also invites comparison with the American market for interns and allows the hypothesis advanced in Roth (1984a) about the behavior of that market to be tested and refined.¹

The particular forms of market organization in the United Kingdom that are the subject of this paper arose in reaction to problems that emerged in the 1960's. Prior to the mid-1960's, the preregistration house-officer markets in the various regions of the National Health Service were largely run in a decentralized way, with students responsible for finding positions on their own and with consultants (as senior physicians and surgeons are called) responsible for filling the positions under their supervision. Competition among students for desirable positions and among consultants for desirable house officers eventually led to

*Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260. In the course of gathering information about the markets described here, I have been helped by numerous British physicians and medical administrators, who have taken the time to correspond with me at length on these matters and sometimes to unearth old records. Among those who have taken pains to help me, I would be remiss not to mention Drs. J. Anderson, T. J. Bayley, P. G. Bevan, K. C. Calman, S. C. Farrow, J. Fraser, F. J. Goodwin, T. M. Hayes, K. Johns, J. H. Lazarus, D. McInnes, G. A. Moge, R. Mulligan, K. M. Parry, R. P. Ryan, D. A. Shaw, D. M. Taylor, H. R. A. Townsend, and N. D. Wright. Of course, they are in no way responsible for any errors or omissions in my account of these markets, nor for the conclusions reached in this paper. I am also grateful to Dr. Susan Mongell, who helped assemble the published literature on this topic. This work has been supported by the National Science Foundation and the Alfred P. Sloan Foundation.

¹Some of my conclusions about this comparison were earlier described briefly in Roth (1990).

these positions being filled earlier and earlier in the students' education. The situation is described as follows by A. Doig and G. Munday (1969):

The filling of preregistration house-officer posts by private arrangements between consultants and students is considered unsatisfactory by most students and by an increasing number of consultants. This practice allows consultants to offer appointments to promising students early in their clinical years and often causes subsequent regret on the part of those who accept when they are later offered more desirable posts; broken agreements cause considerable inconvenience and irritation to consultants. [p. 1250]

J. Alexander-Williams and Ivor Stephenson (1973), writing about the situation in Birmingham prior to the mid-1960's, describe things similarly.

Our graduates who wished to take posts in the teaching hospital insisted on promises earlier and earlier in their undergraduate careers as a result of the many more attractive posts in the region being promised years in advance to graduates imported from other regions. [p. 605]

Doig and Munday further note that

...an initial attempt was made to effect improvement by simply standardising the dates for the receipt of applications and the offering of posts; this procedure had the unfortunate effect of telescoping many of the difficulties and frustrations which had previously occurred over a year or more into a much shorter period. [1969 p. 1250]

This early history is strikingly similar to that of the American market for interns from 1900 to 1951, described in Roth (1984a). By 1944, the standard date at which internship appointments were made in the American market had advanced to two full years before the internship was to begin. In 1945, standardized appointment dates nearer the beginning of employment were

successfully established through the intervention of the medical schools. However, problems manifested themselves in the waiting period between the time offers of internships were made and the time students were required to accept them. Roth (1984a) describes these as follows:

...a student who, was offered an internship at, say, his third choice hospital, and who was informed he was an alternate (i.e., on a waiting list) at his second choice, would be inclined to wait as long as possible before accepting the position he had been offered, in the hope of eventually being offered a preferable position. Students who were pressured into accepting offers before their alternate status was resolved were unhappy if they were ultimately offered a preferable position, and hospitals whose candidates waited until the last minute to reject them were unhappy if their preferred alternate candidates had in the meantime already accepted positions. Hospitals were unhappier still when a candidate who had indicated acceptance subsequently failed to fulfill his commitment after receiving a preferable offer. In response to pressure originating chiefly from the hospitals, a series of small procedural adjustments were made in the years 1945-51. [p. 994]

These adjustments primarily involved shortening the time during which students were allowed to hold offers without either accepting or rejecting them. By 1949, a time of only 12 hours had been rejected as too long, and by 1951 it was widely recognized both that there were serious problems in the last stage of the matching process and that these could not be resolved by compressing this last stage into shorter and shorter time periods.

At this juncture, a centralized (and computerized²) matching procedure was proposed and adopted in the American market.

²Initially the procedure was carried out with card-sorting machines.

Students and hospital programs continued to interview one another as before, but then students submitted to a central clearinghouse a list of hospital programs ranked in order of preference, and hospital programs submitted a list of students. On the appointed day, students were notified of the hospital program to which they had been matched, hospital programs were notified of the students with which they had been matched, and both hospitals and students were encouraged to sign the necessary contracts with one another. Note that at every stage of the procedure, from the decision to submit preference lists to the decision whether to accept the match ultimately proposed, this matching procedure was instituted as a *voluntary* procedure. It is therefore particularly notable that it achieved very high rates of voluntary participation, with the proposed matchings accounting for the large preponderance of jobs filled in this market for many years following its inception. This matching procedure has survived for many years; the algorithm used today to accomplish the match is still essentially the one first used in 1952.

When a Royal Commission on Medical Education (1965–68) issued a report on the problems confronting the preregistration market, the organization of the American market presented an obvious alternative. Many medical schools and their affiliated regional hospitals introduced centralized matching procedures, but different regions used different algorithms to determine the match from the submitted preferences. (Some, but by no means all, of these centralized schemes were implemented by computer.) It appears that more than a dozen regional matching procedures were introduced, but in contrast to the American experience, only a few have survived to the present. Most failed to solve the problems that motivated their introduction and were abandoned.

The principal goal of the present paper is to investigate what common properties might distinguish those centralized procedures that failed and were abandoned from those that succeeded and are still in use today. This will also present an opportunity

to test and refine, in a different environment, the hypothesis advanced in Roth (1984a) about the success of the centralized procedure introduced in the American market.

It was shown in Roth (1984a) that the algorithm employed in the American market produces a *stable* matching in terms of the submitted preferences, where a matching is called stable if no student or hospital is matched to an unacceptable mate and if there is no student and hospital not matched to one another who would both prefer to be matched to each other than to (one of) their partner(s) in the matching. The next section recounts briefly how the major developments in the American market can be explained in terms of stability or the lack of it, which leads naturally to the hypothesis that the success of the American procedure is intimately related to the stability of the matching it produces. Therefore, one goal of the present study is to provide a test of this “stability hypothesis” as it applies to the U.K. markets. Another goal is to consider the incentives agents may have to submit other than their true preferences and the effects this may have.

The regional markets in the United Kingdom provide a stringent test of the stability hypothesis for two reasons. First, administrative authority is more centralized in those markets than in the American market. In some of the regions in which a centralized matching procedure was tried in the United Kingdom, for the period in which it was in effect it was the only way to formalize employment agreements. In contrast to the American market, students in such regions could not decline a job with which they had been matched and arrange instead another job in the same region, and consultants could not decline a student with whom they had been matched in favor of another participant in the match.³ However, as will be seen, it was sometimes possible for students and consultants to circumvent the match by

³However, students could sometimes arrange jobs in other regions.

coordinating with one another before the formal match. The second reason why instabilities might have less consequence in the U.K. markets than in the American market is that the U.K. markets are smaller than the American market by a factor of about 100. Consequently, there are many reasons why, even in regions in which other avenues to formalize employment agreements existed in principle, students and consultants might feel obliged to "play by the rules" and accept a suggested match, even if a better one might have been available. Such reasons might not apply in the much larger and more impersonal American market.

Even in the U.K. markets, the stability of matchings plays an important role, as will be seen below. Perhaps even more important than the formal stability conditions is whether agents can anticipate (or detect *ex post*) and act upon any instabilities that may occur. The centralized markets that were unstable in this way fared no better than the decentralized markets they replaced.

This paper concentrates on seven different procedures by which regional markets in the United Kingdom have been organized. These seven are all of those for which I have been able to obtain sufficiently precise descriptions of the matching algorithm (all seven were computerized). Of these seven, two are based upon modifications of an algorithm that always produces stable matchings, and both of these have controlled the unravelling of appointment dates and survived to the present. The five remaining schemes are based upon algorithms that may frequently produce instabilities. Only two of these have survived (and these are in the two smallest markets); the other three have been abandoned.

The organization of this paper is as follows. Section I discusses the American market. Section II describes the chief differences between the American market and the U.K. markets and describes a formal model suitable for the U.K. markets. The main result of this section is that, even at a stable matching, in these markets there may exist higher-order instabilities of a kind not present in the American market. Section III

begins the analysis of particular market organizations in the United Kingdom, starting with three defunct schemes, which all had particularly severe problems due to the fact that agents had incentives to submit other than their true preferences and to form pairwise coalitions prior to the formal match. Section IV discusses the two markets organized around stable matching procedures and the modifications that have been made in each; and Section V considers the two other surviving schemes and the particular markets in which they operate. Section VI considers some of the difficulties presented by the fact that there is no real national market in the United Kingdom for preregistration positions, but only a collection of regional markets, and conclusions are presented in Section VII. Some of the necessary formal apparatus and proofs will be deferred to the Appendix.⁴ (See Roth and Marilda Sotomayor [1990] for a full account of the literature.)

I. The American Market

In the American market, students each seek one position, while each hospital seeks some number of students.⁵ The size of the market has approximately doubled since 1952: today roughly 20,000 positions are offered annually.

Interns' salaries are part of the job description and are not negotiated by each hospital and intern. Therefore, salaries will not play an explicit role in the model but will simply be one of the factors that determine the preferences that students have over hospitals. Similarly, hospitals can rank students. Because a hospital typically employs more than one student, a full description of a hospital's preferences must include how it

⁴Because of the way the paper is organized, it may be easiest to read the Appendix last, rather than referring to it whenever a proof is deferred to the Appendix.

⁵Strictly speaking, the agents on the institution side of the market are hospital programs rather than hospitals, since different internship programs within a hospital are separately administered and students apply to specific programs.

evaluates alternative groups of students. A simple assumption connecting a hospital's preferences for groups of students to its ranking of individual students is that, if two groups of students differ only by a single individual, the hospital prefers the group containing the higher-ranked individual. Preferences of this sort are said to be *responsive* (to the hospital's ranking of individual students). (More general assumptions are possible, as will be discussed in the next section.) A student is *unacceptable* to a hospital if the hospital would prefer to keep a position vacant rather than fill it with that student, and a hospital is unacceptable to a student if, rather than accept one of its positions, the student would prefer to remain unmatched (and seek employment in a secondary market).

An outcome of the market is a matching of students and hospitals, such that no hospital is assigned more students than it has positions and no student is assigned to more than one position. A matching is *stable* if no student is matched to an unacceptable hospital, no hospital is matched to an unacceptable student, and no student and hospital who are not matched to one another would both prefer to be matched together. (Specifically, the hospital must prefer the student to one of the students it is matched with, or if it has some unfilled positions, it must prefer the student to leaving a position unfilled.)

In Roth (1984a), I undertook to explain three episodes in the history of this market. From 1945 through 1951, the market was characterized by chaotic last-minute recontracting, with students seeking to improve on the positions they had been firmly offered (and had sometimes accepted) by contacting the hospitals they preferred and with hospitals sometimes pressuring students into premature decisions in order to be able to contact students on their waiting lists promptly. This behavior persisted despite repeated attempts by various medical organizations to establish more orderly norms of behavior. However, from 1952, following the introduction of the centralized matching procedure, until the mid-1970's, there was a very high degree of voluntary orderly partic-

ipation, with in the neighborhood of 95 percent of American medical school graduates participating in the matching procedure and ultimately being offered and accepting the position they were matched with. However starting in the mid-1970's, as increasing numbers of married couples sought to obtain two positions in the same vicinity, the rate of participation in the match began to drop, with high percentages of married couples seeking and finding positions outside of the centralized match. So there are two transitions to explain: 1) the transition from chaotic recontracting to orderly voluntary participation that took place in 1952 and 2) the transition from uniformly high rates of participation among medical school graduates prior to the 1970's to the defection of married couples in the late 1970's and early 1980's.

The stability hypothesis is based on the demonstration in Roth (1984a) that the 1952 matching algorithm produces a stable matching (in terms of any preferences that may be submitted) and that the procedure used to assign married couples two jobs in the same vicinity was particularly prone to produce unstable matchings.⁶ Note that a student who has been offered or has had proposed to him a specific job (or a couple which has been matched with a pair of jobs) has only to make a few phone calls to determine whether any of his preferred hospitals would be willing to offer him a position, so the problem of determining whether there are any exploitable instabilities is not a difficult one. Thus, the "stability hypothesis" applied to this market is that the chaotic conditions prior to 1952 reflected the instabilities then present in the market, that the success of the centralized procedure was

⁶In recent years, changes in the way married couples are handled by the match show some signs of ameliorating this problem. However, the problem is not completely tractable, since it was shown in Roth (1984a) that, when there are married couples in the market, the set of stable matchings may be empty. It should also be noted that other developments in the match during the 1970's contributed to the decline in participation during that period.

due to the stability of the matching it produced, and that the decline in participation among married couples in the 1970's was because they once again found instabilities.

Of course, even though the stability hypothesis seems to account for the major developments in this market, the real explanations might lie elsewhere. For example, it might be postulated that *any* centralized market organization would have solved the problems experienced prior to 1951 and that the experience of married couples has less to do with instabilities of the kind dealt with here than with the difficulties young couples have in making decisions. Therefore, it is particularly illuminating to be able to consider the kind of natural experiment, involving both stable and unstable centralized matching mechanisms, that one finds in the United Kingdom.

Before moving on to the U.K. markets, it will be useful to make two somewhat technical observations relevant both to those markets and to the American market. The first is that agents may have incentives to submit rank-orderings to the centralized matching procedure that differ from their true preferences.⁷ It was shown in Roth (1982) that no stable matching mechanism exists that makes it a dominant strategy for all agents to state their true preferences, and it was shown in Roth (1985a) that no stable procedure can make it a dominant strategy for all hospital programs with more than one position to rank students in order of their true preferences. In the American market, stating true preferences is not a dominant strategy for either students or hospitals.

This raises another question about the stability hypothesis: if agents may have reason not to submit their true preferences to the centralized matchmaker, then the fact that the matching algorithm produces a matching that is stable with respect to the

submitted preferences does not assure that the matching is stable with respect to the true preferences (i.e., the preferences according to which agents search for and accept alternative opportunities).

One approach to addressing this question is to consider what will happen when the rank-orderings that agents submit are in equilibrium, even though they may be different from the true preferences. It was shown in Roth (1984b) that, when algorithms equivalent to the one used in the American market are employed, every Nash equilibrium in undominated strategies yields a matching that is stable with respect to the true preferences (as well as to the stated preferences) in the special case arising when all hospital programs have only a single position to fill. In the (actual) case, when hospitals fill multiple positions, fewer strategies are dominated, and so only weaker results have so far been obtained: there are equilibria at which the resulting matching is stable with respect to the true preferences, and other equilibria such that it is not (Roth, 1985a). Another approach is to ask whether the agents in the market have the kind of information about one another's preferences needed to profit from submitting rank-orderings different from their true preferences. (These information requirements are considerable; cf. Roth, 1989.) If not, submitted preferences might approximate true preferences sufficiently to produce stable outcomes. Viewed in this way, the question remains largely an empirical one, which gives further reason to make additional observations of the kind considered in this paper.

The second technical observation concerns the significance of defining stability in terms of individual agents and pairs of agents, without reference to larger coalitions of agents, such as coalitions consisting of a hospital and several students (all of whom might be employable by the same hospital) or coalitions consisting of multiple hospitals and students. When the rules of the market are that any student and hospital may sign a contract with one another if they both agree, the following result from Roth (1985b) says essentially that nothing is

⁷In 1951, an algorithm for the American market was proposed that gave agents clear incentives to state rank-orderings different from their true preferences. For this reason, it was replaced by the 1952 algorithm, which was claimed to give agents no such incentives.

lost by not considering such coalitions explicitly.

PROPOSITION 1: *The set of stable outcomes equals the core defined by weak domination⁸ of the market in which hospitals employ multiple students but students take only one position.*

It will be seen that the conclusions of Proposition 1 do not carry over to the U.K. markets considered next, as a consequence of the fact that most students seek two positions in those markets.

II. The U.K. Markets

A medical school graduate in the United Kingdom is eligible only for provisional registration with the General Medical Council. To become eligible for full registration, a doctor must complete 12 months in a pre-registration position, typically six months in a medical position and six months in a surgical position. These positions are supervised by different consultants and are arranged separately, so each graduating medical student must typically find two pre-registration positions. An outcome of this market is thus a matching of students and consultants such that no consultant is assigned more students than he has positions and no student is assigned more than *two* positions, one medical and one surgical.

Most positions are filled on a regional basis, with graduates of a given medical school going to one of the associated teaching hospitals or to other hospitals in the

same region.⁹ Despite their regional character, these markets have a centralized component absent from the American market; the Department of Health and Social Security sets targets for the number of preregistration house-officer posts for each English Regional Health Authority. These regional markets are two full orders of magnitude smaller than the American market: for 1983 the largest English region was the West Midlands, with just over 300 positions, and the smallest was East Anglia, with just over 100 positions.

The fact that students seek two positions, rather than one as in the American market, means that a different model must be used to represent the market. The simplest modification is to assume that students have separate rank-orderings over medical consultants and over surgical consultants. As in the model of the American market, consultants have rank-orderings over students, but now *both* sides of the market must have preferences not just over individuals but over sets; that is, consultants have preferences over groups of students, and students have preferences over pairs of jobs. Again, the simplest assumption is that these preferences are responsive to the rank-orderings of individuals, as defined in the previous section. As before, a matching is stable if no matches are unacceptable and if no student and consultant who are not matched to one another would both prefer to be matched together. However, even when the assumptions of the model are kept closely comparable to those for the American market in this way, the fact that matchings are many-to-two (i.e., consultants have more than one student, and students need two jobs) rather than many-to-one has the consequence that stable matchings need no longer be in the core of the market or even be Pareto efficient. That is, in this market there is no

⁸An outcome x of a market is *dominated* if there is some coalition S of agents that, by trading among themselves, can obtain allocations they all prefer to x . The outcome x is *weakly dominated* if such a coalition S can obtain allocations that all its members like at least as well as x and that at least one member strictly prefers to x . The core is the set of outcomes that are undominated, and the core defined by weak domination is the subset of the core consisting of outcomes that are not even weakly dominated. Any core outcome is Pareto optimal.

⁹However, in London, where there are many more graduates of local medical schools than local preregistration positions, medical schools commonly have arrangements with hospitals elsewhere.

parallel to Proposition 1. Instead, one has the following result, whose proof is in the Appendix:

PROPOSITION 2: *In the many-to-two matching model with responsive preferences, the set of stable matchings is nonempty for all preferences, but a stable matching need not be in the core and need not even be Pareto optimal.*

The proposition implies that, even when no student and consultant can together arrange to do better than a given matching, there might be a larger coalition, consisting of many consultants and students, who by rearranging job assignments could obtain preferred assignments for all its members. Needless to say, identifying and organizing large coalitions may be more difficult than making private arrangements between two parties, and it will become clear in what follows that the set of stable matchings is still of primary concern.

There are still further generalizations that must be made to the model to allow it to represent faithfully the variety of special constraints found in the various regional markets. In particular, the assumption that agents have responsive preferences must sometimes be relaxed to allow more complex connections between rank-orderings of individuals and preferences over groups.

On the student side of the market, these complexities arise because students require one of each of two types of jobs. However, perhaps the most unusual example of the sort of complicated preferences I have in mind arises on the consultant side of the market administered from Edinburgh. Following what I gather must have been traditional practice¹⁰ before the introduction of a centralized matching scheme, surgeons have the option of indicating that they do not wish to employ more than one female preregistration house officer at a time. Therefore, a consultant surgeon could conceivably submit a rank-order list in which his top four choices, say, were female, but

nevertheless indicate that he wishes to employ at most one of these at any one time.

It is easy to see that these preferences are not responsive to the rank-ordering of individuals as discussed above, since the surgeon in question would prefer to employ his first and fifth choices rather than his first and second choices. Similarly, students clearly prefer one medical and one surgical position to any other combination, regardless of their preferences for individual positions. It will therefore sometimes be convenient to model agents' preferences for sets of alternatives directly, without explicitly considering their rank-orderings of individual students or positions.

Thus, faced with a set S of student applicants, a consultant C can determine which subset of S he would most prefer to hire. We call this C 's choice from S , and denote it by $Ch_C(S)$. That is, for any set S of students, C 's choice set is $Ch_C(S) = S'$ such that S' is contained in S and S' is (at least weakly) preferred to any other subset S'' of S . It will be convenient (but not essential) in what follows to assume that all agents have strict preferences, so that the choice set is unique. Thus, a consultant's preferences are given by a rank-ordering of sets of students, $S_1, S_2, \dots, S_k, \Phi, \dots$, with S_1 being his first-choice set of house officers and so forth (and with any unacceptable set being less preferred than the empty set). These preferences determine the choice function (i.e., for any set S , $Ch_C(S)$ is C 's most preferred subset S' of S). A student's preferences over (pairs of) positions can be represented analogously.¹¹ The constraints mentioned above are consistent with the preferences meeting the following condition:

Definition: An agent A (a consultant or a student) has *substitutable preferences* over sets of alternatives (i.e., sets of students or pairs of jobs) if, for any set T that contains

¹⁰Particularly among urological surgeons.

¹¹In this representation of preferences, the requirement that students have one medical and one surgical position would be represented by having sets consisting of a pair of two medical jobs, for instance, be unacceptable to the student.

distinct elements t and t' , if t is in $\text{Ch}_A(T)$ then t is in $\text{Ch}_A(T - t')$.

If a consultant, for example, has substitutable preferences, then if his preferred set of employees from T includes student t , so will his preferred set of employees from any subset of T that still includes t . (This follows by repeated application of the definition.) Therefore, the consultant regards student t and the other students more as substitutes than as complements and continues to want to employ t even if some of the other students in his choice set become unavailable. Note that responsive preferences are substitutable, so this condition on preferences is a generalization of what has been considered so far.¹² One can similarly generalize the definition of stability as follows.

Consider a matching μ which assigns to each consultant C the set of students $\mu(C)$ and to each student s the set of (no more than two) jobs $\mu(s)$. The matching μ is stable if there is no student s who would prefer to reject one of the jobs in $\mu(s)$, no consultant C who would prefer to reject one of the students in $\mu(C)$, and no student s and consultant C who are not matched to one another but who would prefer to be. That is, μ is stable if there is no student s such that $\mu(s) \neq \text{Ch}_s(\mu(s))$, no consultant C such that $\mu(C) \neq \text{Ch}_C(\mu(C))$, and no student s and consultant C such that s is not in $\mu(C)$ [and so C is not in $\mu(s)$] but such that s is contained in $\text{Ch}_C(\mu(C) \cup s)$ and C is contained in $\text{Ch}_s(\mu(s) \cup C)$. It is possible to state the following proposition, which is proved in the Appendix:

PROPOSITION 3: *In many-to-two matching, when all agents have substitutable preferences, the set of stable matchings is nonempty.*

¹²Substitutable preferences in two-sided matching were first studied by Alexander Kelso and Vincent Crawford (1982). Charles Blair (1988) showed that, in many-to-many matching with substitutable preferences, the core could be empty even though the set of stable matchings is not. (Proposition 2 shows that this is related to many-to-many matching, not merely to complex preferences.)

In summary, although there are clear similarities between the American market for interns and the regional markets in the United Kingdom for preregistration house officers, there are also some important differences. Some of these have to do with the fact that the U.K. regional markets are both more centralized and much smaller than the American market. Other differences have to do with the fact that students in the U.K. markets seek two jobs. As a result of the latter differences, preferences in the U.K. markets cannot always be modeled as simply as in the American market and require the more general model described in this section.

In considering the operation of the various regional markets, the more general model will be required to establish what is a stable matching. However, when the procedures that produce unstable matchings are considered, it will be possible to demonstrate the instabilities even within the confines of the simpler models.

III. Matching by Priority: Newcastle, Birmingham, and Edinburgh (1967)

This section considers three closely related matching schemes, all developed in the late 1960's, and all subsequently abandoned.¹³ In each of these schemes, a student's ranking of a particular consultant was combined with the consultant's ranking of that student to produce a "priority" for that student to be employed by that consultant. The three schemes differ in the way in which this priority was determined, and each will be discussed below with examples. In each scheme, the overall matching of students with consultants was determined by making the individual matches of students with consultants in order of priority. That is, the first step of each of the three algorithms was to make all of the first-priority matches. Then consultants with unfilled positions and students still needing jobs were scanned to

¹³The Edinburgh scheme was replaced around 1969 by another centralized system, considered in the next section.

identify any second-priority matches, and so on.

The schemes introduced in Newcastle in 1967 (A. G. Leishman and R. P. Ryan, 1970) and in Birmingham in 1966 (Alexander-Williams and Stephenson, 1973) were almost identical. They each used the product of the student's ranking of the consultant and the consultant's ranking of the student as the basis for the priorities. If a consultant and student each ranked one another first [a "(1,1) match"], they had a priority of 1. If the consultant ranked the student first but the student ranked the consultant second [a "(1,2) match"], they had a priority of 2, as did a consultant who ranked a student second but was ranked first by the student [a "(2,1) match"]. The two schemes differed in how they broke ties: in Birmingham, ties were broken in the consultant's favor, so that a (1,2) match would have a higher priority than a (2,1) match. In Newcastle, ties were broken in the student's favor.¹⁴

In the scheme introduced in Edinburgh in 1967, priorities were lexicographic in consultants' preferences. That is, (1,1) matches were the first priority, followed by (1,2), (1,3), (1,4), and so forth. Only when all consultants' first choices had been exhausted were other matches [(2,1), (2,2), (2,3), etc.] considered. The following example will illustrate the similarities and differences among these three schemes and also prove the following proposition about them:

PROPOSITION 4: *Each of these schemes may produce unstable matchings.*¹⁵

Example 1. For simplicity, consider six consultants, each of whom has only one position to fill, and six students, each of whom

needs only one position. (It will be clear that the example does not depend on this simplification.) The rank-orderings of the agents are as follows.

$C_1: s_1, \dots$	$s_1: C_1, \dots$
$C_2: s_1, s_3, s_2, s_4, s_5, s_6$	$s_2: C_2, C_1, C_3, C_4, C_5, C_6$
$C_3: s_3, s_4, \dots$	$s_3: C_4, C_3, \dots$
$C_4: s_4, s_3, \dots$	$s_4: C_3, C_4, \dots$
$C_5: s_1, s_2, s_5, s_3, s_4, s_6$	$s_5: C_1, C_2, C_5, C_3, C_4, C_6$
$C_6: s_2, s_5, \dots$	$s_6: C_5, C_2, \dots$

Then the Birmingham algorithm makes the following matches (the priority is indicated in parentheses after each set of matches): C_1s_1 (1,1), C_3s_3 and C_4s_4 (1,2), C_2s_2 (3,1), C_5s_6 (6,1), C_6s_5 (2,6). This outcome is unstable because C_5 and s_5 are one another's third choices, but in the Birmingham match they are not matched to each other, but are each matched to their sixth choices.

The Newcastle algorithm makes the matches: C_1s_1 (1,1), C_3s_4 and C_4s_3 (2,1), C_2s_2 (3,1), C_5s_6 (6,1), C_6s_5 (2,6). This outcome is also unstable with respect to C_5 and s_5 .

The Edinburgh (1967) algorithm makes the following matches: C_1s_1 (1,1), C_3s_3 and C_4s_4 (1,2), C_6s_2 (1,6), C_5s_5 (3,3), C_2s_6 (6,2). This outcome is also unstable, but with respect to C_2 and s_2 , who would each prefer one another to their assigned partners (they are each matched to their sixth choices).

So far, the example has been analyzed as if the agents all state their true preferences. Before considering the incentives that agents may have to do otherwise, it will be illuminating to examine the history of these matching systems after their introduction and how they failed and were abandoned.

The various accounts I have received of the demise of these systems all agree on the main events.¹⁶ The following description is

¹⁴At least initially. A later modification was to reverse this method of tie-breaking. I am indebted to Dr. D. A. Shaw, the Dean of Medicine at Newcastle, for this observation (pers. comm., 13 May 1987).

¹⁵A stronger result, namely that *any* priority matching scheme will sometimes produce unstable matchings, is proved in Proposition 10 in the Appendix.

¹⁶I have been able to obtain the least information about the demise of the Edinburgh system, which by 1969 had already been replaced. I have been able to obtain much more detailed accounts, from multiple sources, of the experiences at Newcastle and Birmingham.

from a letter by Dr. John Anderson, the Postgraduate Dean at Newcastle (pers. comm., 6 May 1987):

To understand why our computerized scheme was discarded [in 1981], you should know that in the Northern Region there are 202 recognised posts (this target is set by the DHSS [Department of Health and Social Security]) in 26 approved hospitals. Each year we normally graduate a maximum of 130 students, so that we regularly have a shortfall of at least 70 and usually more, since a number of our graduates obtain pre-registration posts in other regions. We are therefore a major importing region and each six months fill between 50 and 60 posts with graduates of other regions. However, we have never filled more than 185 posts and this means that up to 20 pre-registration posts are regularly unfilled. Sometimes Senior House Officers will be appointed to these posts, but every six months there is a small number of posts that are left unfilled.

This is the background to our problems, and this imbalance between local graduates and posts explains why the computerized scheme failed. Understandably, consultants in the periphery of the region were anxious to fill their posts as quickly as possible and often entered into private arrangements with undergraduates. ...the practice of making private arrangements outwith the computer match scheme gradually spread to the Teaching Hospitals. Those who stuck rigidly to the scheme often found that they were left without any housemen to appoint, as there was no way of preventing these private arrangements and no sanctions could be introduced against those who operated outside the scheme.

In the late 70s and early 80s an increasing number of problems cropped up, mainly concerning conflicts between private arrangements and the formal application procedure. There was a feeling that the computer scheme was an impersonal mechanism which inhibited personal contact between students and consultants and

shortly before the scheme was discarded we found that in up to 80% of cases students and consultants only used the computer to indicate a first preference.... The main reason for the abandonment of the scheme, therefore, was that there were problems in getting students and consultants to participate in an orderly way, and this led to those who rigidly observed the requirements of the scheme to be penalized.

The experience in Birmingham was similar, but there the centralized procedure, which was initiated in 1966 for a limited group of hospitals, failed after a few years, was resumed on a larger scale in 1971, failed once more, was restarted again around 1978, and was finally abandoned again around 1981. Of the initial experience, Alexander-Williams and Stephenson (1973) say:

Perhaps the most important cause of failure was the lack of enthusiasm by the consultant staff who did not wish to be deprived of the right to choose their junior staff and so, suspicious that the matching programme might allocate them someone whom they did not want, still tended to make a promise to the first acceptable student who approached them. While there are obviously no objections to first choices being mutually agreed, it soon became known among the undergraduates that certain posts were already promised and so began once again the unseemly struggle that the matching programme was designed to avoid. The breakdown of the scheme was to the disadvantage of the diffident student or the one who waited until he had 'surveyed the field.' [p. 606]

Dr. P. G. Bevan, the director of the board of graduate clinical studies at the University of Birmingham Medical School, writes as follows about the most recent attempt (pers. comm., 2 November 1984):

The main cause of failure for this Matching Plan was the fact that too many Consultants did not abide by the conditions and promised their job to

one particular Student.... In view of this we finally abandoned the Matching Plan three years ago.

A recurrent theme in these accounts is that, after the centralized matching had been in use for a short while, increasing numbers of jobs began once again to be privately arranged in advance between consultants and students and that this worked to the detriment of those who tried to participate in the scheme without prior arrangement.

To understand this phenomenon, consider now the incentives which these priority procedures give to the agents. For this purpose, consider again Example 1. (In the example, there are equal numbers of students and positions, but it will be clear that the behavior described below could exist at least as easily when there is an imbalance between the two.) To make the example clear, suppose consultants C_1 – C_4 are in the most desirable teaching hospital, C_5 is in the next most desirable regional hospital, and C_6 is in a relatively undesirable rural hospital. Similarly, suppose students s_1 – s_5 are all top graduates of their medical school, while s_6 has a less distinguished record.

Then, under the Birmingham or Newcastle system, C_5 is gravely disappointed to learn that his new junior house officer will be s_6 , all the more when he learns that student s_5 , whom he liked reasonably well, is quite unhappy with his own appointment and would have preferred to work for C_5 . If C_5 had submitted a rank-ordering on which s_5 was his first choice, they would have been matched, as would also have been the case if s_5 had submitted a rank-ordering on which C_5 was his first choice. The example shows there may be incentives for both students and consultants to submit rank-orderings different from their true preferences.

Furthermore, these priority-ranking systems allow such incentives to build upon one another, so that as more agents adapt their submitted rank-orderings to improve their matches, the greater is the incentive for other agents to do so. To see this, suppose C_5 in Example 1 resolves not to suffer the same fate the following year. He there-

fore approaches one of the good students in the next year's class, in advance of the formal match, and suggests that they mutually agree to be matched, which they will accomplish by ranking one another first in the formal match.¹⁷ The student, chastened by the experience of s_5 the previous year, is receptive. Now consider the situation in the formal match, when a number of positions have been prearranged to be (1,1) matches. Suppose students t_1 , t_2 , and t_3 have made such arrangements with consultants C_3 , C_4 , and C_5 , but consultant C_2 , not knowing this, submits his true rank-ordering, which is $t_1, t_2, t_3, t_4, t_5, \dots$, and t_4 submits his true rank-ordering $C_3, C_4, C_5, C_2, \dots$. Although C_2 does not know it, t_4 is his highest-ranking student who is actually available, and C_2 is t_4 's most-preferred available consultant. However, since the product of their rankings is 16, C_2 could well end up with his 15th-choice student. So when some matches have been prearranged, those not in the know, students as well as consultants, stand to do very poorly if they do not also prearrange their matches. Furthermore, when an agent's top n choices have all arranged to indicate only a first preference in the formal match (as in the above quote from Anderson), then the agent can do no better in the match than to reach such an agreement himself with his $n+1$ st choice. Therefore, the following proposition, which applies as well to the Edinburgh (1967) system,¹⁸ has been proved:

PROPOSITION 5: *It is not a dominant strategy for any agent to submit his true preferences in these priority matching systems.*

¹⁷In the context of these relatively small markets, both parties to such an agreement can be confident that it will be carried out; since a consultant with a reputation for not delivering on his promises will soon find it difficult to attract good junior house officers, and a junior physician is reluctant to incur the enmity of a senior physician in the region in which he hopes to practice.

¹⁸Under the Edinburgh (1967) system, C_2 in Example 1 could have improved his match by ranking s_2 first, and s_2 could have improved his match by ranking C_6 as unacceptable.

Furthermore, there are multiple equilibria at which all agents submit only a first choice.

In fact, Proposition 5 does not capture the full strength of what has been proved. Under priority matching, a student and consultant who rank one another first will be matched regardless of what the other agents do. So the problems of coordination that afflict many equilibria do not arise here: pairs of agents may secure their part of the equilibrium by private arrangement. These results thus go a long way toward explaining both why a high percentage of appointments were soon arranged in advance under these systems and why this worked to the disadvantage of those who tried to arrange employment through these priority-based formal match procedures.¹⁹

IV. Stable Matching: Edinburgh (1969) and Cardiff

This section considers two very closely related systems, both built around the same stable matching algorithm. That the two systems are closely related is not an accident: the Cardiff system was adapted around 1971 from the system initiated in Edinburgh around 1969 to replace the Edinburgh (1967) priority matching system. However the two markets are rather different, with Cardiff regularly having more positions than local graduates and with Edinburgh regularly having more graduates than positions.²⁰ Both systems remain in operation today,

having achieved and maintained high rates of orderly participation.

It is also not an accident that the basic algorithm produces stable matchings, but it is a curious bit of intellectual history. In the 1960's and 1970's the architects of the various matching systems in the United Kingdom knew of the experience of the American markets since 1952, but it was not then known that the American algorithm produced a stable matching. However, unconnected with any of these markets, David Gale and Lloyd Shapley (1962) formulated a simple model of one-to-one matching and defined the set of stable matchings, which in that model equals the core of the game. They also developed a "deferred-acceptance" algorithm (described below) for producing stable matchings and observed that it could be applied as well to problems of many-to-one matching.²¹ Two British computer scientists, D. G. McVitie and L. B. Wilson further explored this algorithm, and apparently through their work, the algorithm came to the attention of Dr. H. R. A. Townsend, the author of the Edinburgh (1969) matching procedure, known as PRAMS (for Pre-Registration Appointment Matching Scheme).²² He used Gale and Shapley's deferred-acceptance algorithm as the basis for PRAMS, adapting it to the requirements of the preregistration market and to the local conditions in Edinburgh. The Edinburgh PRAMS algorithm was subsequently adopted in Cardiff, where it was

¹⁹It appears that something very similar (if not identical) to a priority matching scheme was tried and subsequently abandoned at Sheffield, but I have not included that system among those formally analyzed here because the match was done by a committee, whose exact procedures therefore cannot be determined. However, A. D. Clayden and James Parkhouse (1971 p. 9) report a computer program designed in their words "to mimic the manual allocation," and that program implements a priority algorithm like the Edinburgh (1967) system, except in giving lexicographic priority to students' preferences rather than to consultants' preferences.

²⁰The Edinburgh system is furthermore open to students from medical schools in other regions (Sir James Fraser, pers. comm., 7 May 1987).

²¹Gale and Shapley (1962) concentrated primarily on the one-to-one matching problem, and for many years thereafter one-to-one and many-to-one matching came to be regarded as essentially equivalent. That they are not, in ways that are important for the subject at hand, was first observed in Roth (1985a).

²²Townsend (1981) cites McVitie and Wilson (1971) in the PRAMS.80 manual, which documents the current (as of 1984) version of his computer program, and which he graciously sent to me (pers. comm., 23 November 1984; see also McVitie and Wilson, 1970a, b). Townsend himself is a clinical neurophysiologist, who at the time also held a part-time appointment in Edinburgh's department of Machine Intelligence and Perception, and it was in that capacity that he undertook the design of the PRAMS system (pers. comm., 27 July 1989).

further adapted to local conditions. Before discussing the adaptations introduced in Edinburgh and Cardiff, I will consider first the unmodified²³ deferred-acceptance algorithm for many-to-one matching.

Step 1: Each consultant C with q_C positions makes offers to his q_C highest-ranked acceptable students (or to all of them if there are fewer than q_C).

...

Step k : (i) Each student s rejects all but the highest-ranked of the acceptable offers he has received in steps 1 through $k-1$. (ii) Each consultant C who has received $r \geq 1$ rejections in part (i) of step k , and who now has $q_C - r$ (nonrejected) offers outstanding, makes offers to his $q_C - r$ highest-ranked acceptable students among those who have not yet rejected him. (iii) The algorithm stops at any step $k = T$ at which no rejections are issued, and the resulting matching places each student with the consultant (if any) whose offer he has not rejected and leaves unmatched any student not holding an offer.

Since each student holds at most one unrejected offer at any step of the algorithm and since no consultant makes an offer twice to the same student, the algorithm stops and produces a matching. This matching is stable, because if some consultant C would prefer a different group of students than he receives, then if he has responsive preferences²⁴ he will have proposed to those stu-

dents at an earlier step of the algorithm and been rejected. This implies that these students prefer the positions they get from the algorithm to C , so the matching is stable.

Note that there is another version of the algorithm in which students make applications for positions and each consultant refuses all but the best q_C acceptable offers from among those he has received. While the Edinburgh matching scheme adapts the consultant-proposing version of the algorithm, the present Cardiff version adapts the student-proposing version.²⁵ Both versions produce stable matchings in many-to-one markets.

The markets to which this algorithm was adapted involve many-to-two matchings, and as has already been seen from Proposition 2, this changes the market in important ways. Furthermore, some of the adaptations in both Edinburgh and Cardiff imply that the preferences of the agents are not responsive. Consider the following constraints:

- 1) Each student desires no more than one medical and one surgical position.
- 2) Edinburgh surgeons may specify that they will employ no more than one female house officer in any six-month period.
- 3) In Cardiff (for some of the period under consideration, but not presently), at most one of a student's two positions could be a teaching-hospital position.²⁶

²³Except to allow that some students and consultants may find some matches unacceptable.

²⁴Responsive preferences, which play an essential role in the argument for the case of many-to-one matching, were introduced in Roth (1985a). Earlier treatments of many-to-one matching had argued from analogy with the case of one-to-one matching and had not considered that C 's preferences must be based on comparisons of groups of students. A notable exception is the paper of Kelso and Crawford (1982), which considers preferences defined directly over groups, without reference to an underlying preference over individuals.

²⁵Over the years, the computer code used in Cardiff has undergone a number of programming and procedural changes designed to cope with increasing numbers of positions and with changes in available computers. The current system goes under the name of PASHA (Preferential Allocation System for House Appointments), and during 1973–1982 it went under the name of CHAMP (Computerized House Appointments Matching Plan), during which time the consultant-proposing version of the algorithm was implemented. I am indebted to Dr. Kelvin Johns for documentation of the various systems (pers. comm., 6 November 1984 and 23 June 1989).

²⁶These descriptions abstract somewhat from the actual adaptations. In Edinburgh, a surgeon may actually only specify that he does not wish to employ more than two female house officers, since his positions for two consecutive six-month periods are allocated simul-

Constraints 1 and 2 obviously concern agents' preferences (i.e., certain sets are unacceptable), and constraint 3 can be modeled as part of students' preferences, since it has the effect that students will not (because they cannot) choose two teaching-hospital positions, either during the formal match or in any postmatch exploration of potential instabilities. The Edinburgh adaptation of the algorithm to constraints 1 and 2 is straightforward: students may hold up to two offers at any step in the algorithm but no more than one medical and one surgical, and they must reject the rest. In the Cardiff adaptation to constraints 1 and 3, students could apply at any step of the algorithm to no more than one medical or surgical position and to no more than one teaching-hospital position.²⁷ Subject to these constraints, students' and consultants' offers and rejections within the algorithms are otherwise governed by their submitted rank-orderings of individual positions, as in the deferred-acceptance algorithm described above. (In Edinburgh, students submit two preference lists, one for surgical positions and one for medical. In Cardiff, they submit one preference list containing both kinds of positions.²⁸

taneously. It then remains to schedule the house officers so that two female house officers are not assigned to the same six-month period. Scheduling may present other difficulties as well, some of which may involve "higher-order" instabilities, which will be briefly discussed later. In Cardiff, when constraint 3 was in effect, if some teaching-hospital positions were left unfilled by the initial run of the algorithm, these positions were then "unmarked," so that a student who already had one teaching-hospital position became eligible to fill them.

²⁷This might involve, for example, a student applying at the first step of the algorithm to his first-choice medical position in the teaching hospital and to his fifth-choice surgical position. If he were subsequently rejected by the medical position, he might wish to apply to a surgical position in the teaching hospital, which might be his second-choice position. The algorithm gave him the opportunity to do so, which involved withdrawing his application from his fifth-choice surgical position. Thus, unlike the original deferred-acceptance procedure, applications could be withdrawn as well as rejected.

²⁸This allows medical and surgical teaching-hospital positions to be compared, as they must be to imple-

PROPOSITION 6: *Student preferences satisfying constraint 1 and consultant preferences satisfying constraint 2, but otherwise responsive to a simple rank-ordering, are substitutable. Student preferences satisfying constraints 1 and 3 and otherwise responsive to a simple rank-ordering need not be substitutable.*

Together with Proposition 3, Proposition 6 (which is proved in the Appendix) establishes that stable matchings continue to exist in the Edinburgh market and in the current Cardiff market. The following proposition (also proved in the Appendix) states that the algorithms adapted as described for the Edinburgh market and for the present Cardiff market (without constraint 3) in fact produce stable matchings:

PROPOSITION 7: *The consultant-proposing deferred-acceptance algorithm adapted to preferences that obey constraints 1 and 2 but are otherwise responsive to a simple rank-ordering produces a stable matching, as does the student-proposing algorithm adapted to preferences that obey constraint 1.*

There are some complications that have so far been passed over, the chief of which involves scheduling the jobs each student has been assigned into the August and February starting periods in a way that gives each consultant the right number of house officers in each period. Occasionally (although apparently rarely) the necessary scheduling may be infeasible, as when a number of individuals require a job at a particular time, as a consequence of having arranged their other assignment outside of the region. These situations are resolved at the discretion of the PRAMS committee, typically by editing a job from the preference lists of one of the students (Townsend, 1981 p. 42). Of course this may produce an instability, but it is not one that could be

ment condition 3. Prior to 1973, winter positions in Cardiff were matched before summer positions, but since 1973 they have been matched simultaneously.

predicted in advance, nor can it be acted on following the match, given the rules by which jobs are formally assigned. Although students are invited to indicate the order in which they prefer to fill their medical and surgical positions, these preferences are not honored if they interfere with a feasible scheduling, so there may be some instabilities involving exchange of time slots among groups of students and consultants. I call these instabilities "higher-order" to indicate that they involve coalitions larger than a student and consultant. In the case of scheduling changes, they involve at least two students and two consultants. As noted in Proposition 2, such higher-order instabilities may result from other causes as well, but such higher-order instabilities are likely to be of lesser importance, since they are so difficult to act on. A case in point is the allocation of married couples, which poses the same theoretical difficulties in the United Kingdom as in the United States, but which seems to have had less practical consequence. Whereas an American couple needs to find two jobs in the same location, in the United Kingdom a couple may need to find four jobs in two locations (in two time periods), and the difficulties of identifying a set of such jobs that both the couple and the relevant consultants prefer to those allocated by the match are formidable.

The situation in Cardiff was even more complex during the years in which students were constrained to hold no more than one teaching-hospital position. Aside from the higher-order instabilities just discussed, the combination of constraint 3 with constraint 1 on the preferences may sometimes have the consequence that no stable matchings exist (see Example 4 in the Appendix). In such circumstances, as well as in certain others²⁹ the Cardiff scheme would produce matchings with some instabilities. As near as I can determine (and I do not know how

to make this precise), these instabilities arose relatively rarely.³⁰ What is certain is that they could not be predicted in advance, and so neither the Cardiff system (with or without constraint 3) nor the Edinburgh system allows mutually beneficial "prematch" agreements of the kind discussed in the previous section.³¹

V. Matching by Linear Programming: Cambridge and The London Hospital Medical College

This section considers two closely related matching schemes which do not produce stable matchings in terms of the stated preferences but which are still in use. The first was developed in 1973 at the London Hospital and its Medical College, and the second was developed in 1978 at the University of Cambridge School of Clinical Medicine. Both schemes involve the linear-programming assignment algorithm.

The London Hospital scheme takes as input the rank-orderings of students and consultants.³² Numerical weights are assigned to choices. These are summed for each potential student-consultant pair.

³⁰ I have no information suggesting that constraint 3 was discarded *because* of its potential to produce instabilities; apparently, the facility for automatically accommodating married couples was also eliminated at around the same time, for reasons of simplicity (K. Johns, pers. comm., 23 June 1989).

³¹ Note, however, that in these systems also, students and consultants may assure themselves of being matched by ranking one another first. In this connection, a noteworthy feature of the Cardiff system is that consultants may make their preference list publicly available before the students submit their own preference lists, and in recent years most of them have apparently done so (K. Johns, pers. comm., 3 August 1989). Thus, students will often know where they are in the consultants' rankings before having to submit their own.

³² The students are all graduates of the London Hospital Medical College, and the consultants include all those offering house-officer posts at the London Hospital, together with a group of affiliated hospitals. In a recent year there were approximately 40 posts in London and 70 in regional hospitals. I am indebted to Dr. F. J. Goodwin, the Postgraduate Sub-Dean at the London Hospital Medical College for this information (pers. commun., 31 October, 1984).

²⁹ The full computer code for Cardiff PASHA (as of 1984, including the code for handling constraint 3) is more than 2,000 lines long and creates many special situations, such as those involving filled and unfilled teaching-hospital positions.

A. R. Shah and S. C. Farrow (1976) report that for the first five uses of the algorithm (once in 1973 and twice in 1974 and in 1975) choices 1, 2, 3, and 4 were given weights of 20, 14, 9, and 5, respectively. Thus (1,1) matches received a weight of 40, (1,2) and (2,1) matches each received a weight of 34, and so forth. The resulting weights for each potential student–consultant pair form the basis for the linear-programming assignment problem of matching students to consultants so as to maximize the sum of the matches. Shah and Farrow report that:

The general procedure is for the computer solution to be submitted to the sub-committee of the academic board. On each occasion it has formed a very satisfactory basis for the final solution, although several hand adjustments have been made.³³ [p. 477]

The Cambridge scheme was first implemented in 1978 for posts beginning in February and August 1979. (Students, who normally graduate in December, may apply for positions beginning in February, in August, or for one of each.) Students rank consultants as A, B, C, or unacceptable. "Students may make only two A choices for medicine, and only one of these may be for a medical job at Addenbrooke's, the main teaching hospital. Similarly for surgery. As many B and C choices as they wish are allowed" (D. M. Taylor, pers. comm., 5 June 1987). Consultants rank students similarly (but after learning how the students ranked them, and without a constraint on the number of A rankings.) As in the London scheme, weights are given to each potential match, but in the Cambridge scheme these weights are lexicographic in consultants'

preferences, so consultant–student (A,A) matches have the highest weight, followed by (A,B), (A,C), (B,A), and so forth. As in the London scheme, the matching giving the highest total weight forms the basis of the matching of students and consultants. This matching is also subject to some adjustment (e.g., students are permitted to have only one of their two jobs at Addenbrooke's Hospital). Example 2 proves the following proposition:

PROPOSITION 8: *Both of these schemes may produce unstable matchings.³⁴ Furthermore, they may fail to make (1,1) matches.*

Example 2. For simplicity, consider three consultants, each of whom has only one position to fill, and three students, each of whom needs only one position. Their rank orderings are as follows:

$C_1: s_1, s_2, s_3$	$s_1: C_1, C_2, C_3$
$C_2: s_1, s_3, s_2$	$s_2: C_1, C_3, C_2$
$C_3: s_3, s_1, s_2$	$s_3: C_3, C_2, C_1$

The unique stable matching is μ such that $\mu(C_1) = s_1$, $\mu(C_2) = s_2$, and $\mu(C_3) = s_3$. [This follows since (C_1, s_1) and (C_3, s_3) are both (1,1) matches and so must be included in any stable matching.] However, the London Hospital scheme (with weights as given above) gives μ a weight of 98, while the highest total weight (108) is achieved by the unstable matching ν with $\nu(C_1) = s_2$, $\nu(C_2) = s_1$, and $\nu(C_3) = s_3$. Similarly the Cambridge scheme [with (A,A) matches worth 8, (A,B) worth 7, and so on, down to (C,C) worth 0 for this example] chooses ν (which has a total weight of 20, compared to μ with total weight of 16).

In this example, no individual agent can improve his outcome by changing his stated preference.³⁵ More to the point, even when

³³Shah and Farrow (1976 p. 477) also note that, to adjust the match, the preferences of different individuals may be weighted differently. "[I]n July 1974 the initial solution led to several applicants, who had completed 6 months, not being allocated a second appointment. This made it necessary to rerun the programme with a reduction in weight of newly qualified applicants. This second run achieved the designed outcome."

³⁴That the London scheme may produce unstable matchings was noted by Townsend (1977).

³⁵Actually, this depends on how unacceptable matches are weighted. When unacceptable matches are part of the matching with the highest total weight, they correspond to unmatched students and positions.

all pairs of agents but one organize themselves into (1, 1) matches (as in the example), they cannot be sure of being matched (unless they both rank all other options as unacceptable). Thus, private arrangements are more difficult to make and are less certain than in any of the other matching schemes considered here. Nevertheless, the following parallel to Proposition 5 for priority matching systems applies as well to these linear-programming systems; the proof is left to the reader:

PROPOSITION 9: *It is not a dominant strategy for any agent to submit his true preferences in these linear-programming matching systems. Furthermore, there are multiple equilibria at which all agents submit only a first choice.*

While these systems have some important differences from the failed priority matching systems [as exemplified by the fact that (1, 1) matches are not assured of forming], they have enough significant similarities so that one may wonder how to account for the longevity of the London Hospital and Cambridge schemes. One hypothesis is that the environments in which the markets are conducted differ significantly from other environments: each of these two schemes is for the graduates of a single medical school, on the small end of the range of markets considered here.³⁶ Thus, there may be social and other kinds of pressures that make it difficult to circumvent the formal matching scheme. In this regard S. C. Farrow (pers. comm., 23 October 1984) writes:

Candidates are not obliged to accept allocated posts and have been known to decline; this is of course frowned upon by the scheme and they try hard

to dissuade people from taking such a course.

Similarly, Goodwin (pers. comm., 31 October 1984) adds:

Virtually all candidates accept the posts allocated to them. Indeed, we achieve a match of the candidates' and consultants' first choices in around 70 percent of cases which is obviously satisfactory to all concerned. The remainder nearly always accept the post to which they have been allocated. Of course, we have no legal right to prevent someone from declining a post after it has been allocated to him but this would be considered pretty bad form and the candidates know it.

Thus, in markets of this size and composition, it may be that participation in the matching scheme may be somewhat less voluntary than in larger markets or markets in which many of the agents are not so intimately connected with one another.³⁷

Another hypothesis, in view of the high reported percentage of (1, 1) matches, is that the agents manage to adapt to the system by coordinating among themselves before the formal match or by modifying the rank-orderings they submit. In such a case, the outcome of the match might even be stable. (To test such a hypothesis, it would be necessary to have better information about the submitted rank-orderings than I presently have. However, see Susan Mongell and Roth [1991] for an account of an unstable match-

³⁶They are thus not completely "impersonal" markets, particularly since the most desirable positions are in the associated teaching hospital; and the most desirable house officers are the top students in the class. Since total class sizes are around 100, the key players on both sides of the market have often had the opportunity to get to know one another in the course of the students' medical education.

³⁷B. T. Colvin, the present Postgraduate Medical Sub-Dean at The London Hospital Medical College, writes that "Much of the goodwill in the system relies on the Postgraduate Sub-Dean's personal knowledge of the candidates, consultants and posts, and his ability to impose on both parties the moral obligation to comply with the allocations..." (pers. comm., 10 August 1989). To the extent that agents in this market are obliged or feel obliged to comply, it would of course be unsurprising that unstable matchings and matching procedures can persist: the National Football League draft and the process by which graduates of the U.S. Naval Academy obtain their first assignments (see Roth and Marilda Sotomayor, 1990) are good examples of such markets.

ing procedure for which such data were available; the data gathered for that procedure supported an equilibrium misrepresentation hypothesis of this kind.)

In either case, these matching schemes and the environment in which they operate appear to make prematch coordination both more difficult and less rewarding than do the priority matching schemes discussed in Section III.

VI. Interregional Instabilities and the Current State of Affairs

Students who fail to obtain two positions in the regional markets can participate in a secondary market, called the Safety Net, organized by the Councils for Postgraduate Medical Education,³⁸ which distributes information on unfilled posts and unplaced students. Consultants with unfilled positions may hire senior house officers (instead of preregistration house officers) after a certain date. In 1983, about 25 graduates of U.K. medical schools failed to obtain any preregistration position, in part because some consultants apparently hired senior house officers before the preregistration market had cleared. The incentives for consultants to do so were apparently heightened by the fact that some students would break arrangements made in this market if a more desirable position became available at the last minute (see Henry Yellowlees, 1983; E. D. Acheson, 1984).

In response to these events, a working group was established to study the current market, and their 1987 report (Department of Health and Social Security, 1987) describes the current state of affairs as follows:

...matching arrangements are tending to take place increasingly early in the

clinical years. In some cases, students are matched to posts even before they have entered the final clinical year.... The earlier a student is matched to a post, the greater the chance that for some reason he or she may decide not to fulfil the arrangements but to take some other post instead. Where medical schools run particularly early matching schemes it appears that some Districts not involved in the schemes then advertise their pre-registration posts at an even earlier date in order to try to pre-empt the matching scheme. This has led to some posts being advertised as much as 18 months before their start date. [p. 11]

The report also notes that medical schools should help in:

...ensuring that their unmatched students are made aware that it is unethical, while holding an offer for one post, to accept an offer for another before negotiating release from the prior bargain.... [p. 11]

What seems to be occurring is that, as the centralized matching schemes in individual regions of the National Health Service have unravelled backward in time and been abandoned, the situation in most regions today has come to resemble that of the late 1960's.³⁹ The instabilities that underlie the contemporary problems appear to involve not only instabilities in a given region, but also instabilities between regions.⁴⁰

³⁸The Council for Postgraduate Medical Education in England and Wales has recently been replaced by the Standing Committee on Postgraduate Medical Education (SCOPME). As of 1 April 1989, the responsibility for the Safety Net has passed to the Department of Health (no longer DHSS, social security having been separated from health in mid-1988).

³⁹In another contemporary report, J. H. Gillard and T. H. S. Dent (1988 p. 344) note that "Both matching schemes and free markets with an official start date were reported to be vulnerable to pre-empting. This criticism was conveyed most vigorously from Nottingham and Southampton [both with "free market" decentralized systems], and caused considerable anxiety. Students feared that their colleagues were making private arrangements with consultants, while consultants were keen to avoid missing the more attractive students. If a formal match operates later, students might waste an application if the posts have been covertly allocated by student and consultant agreeing to place each other first in their order of preference."

⁴⁰Instabilities involving positions in other regions can affect even regions with stable regional matching

In view of the experience in the regional markets and in the American market, perhaps the most promising plan to remedy the unraveling of appointment dates and the problems that accompany it would be to replace the patchwork of regional markets with a national market organized around a set of stable procedures (such as those in use in Edinburgh).⁴¹

VII. The Implications for Other Markets

The environment in which the various U.K. markets operate differs in important ways from that of the American market in size, in the number of jobs sought by each student, and in the centralization of authority (at least within a given region). Still, the experience of the various centralized matching schemes considered here allows two very different hypotheses about matching markets which might have been formed on the basis of the evidence of the American market to be rejected.

The first of these hypotheses, which might be called the "pure-transaction-cost hypothesis," is that, because a centralized matching mechanism reduces the high transaction costs found in the chaotic decentralized markets, *any* centralized procedure would obtain high rates of participation. The rapid failure of the priority matching schemes in the regions where they were tried is clear evidence against this hypothesis.

The second of these hypotheses, which might be called the "pure-efficiency hypothesis" (or perhaps the "core hypothesis") is that, in order to achieve high rates of participation, a scheme must produce matchings that do not allow coalitions of *any* size to profit by defecting from the scheme. The

higher-order instabilities that can occur in the Edinburgh (1969) and Cardiff schemes (and which Proposition 2 suggests may be endemic), not to mention the pairwise instabilities that can arise in the linear programming schemes, suggest that this is not the case either.

A more accurate description seems to lie somewhere in between. Centralized matching schemes like the priority matching schemes, which make it easy for pairwise coalitions to defect profitably, experience widespread defection even when the system is endowed with enough authority to require at least pro forma participation. However, orderly participation in a centralized matching scheme appears to be much more likely when the opportunities it presents for defection are rare and difficult to find or when they primarily involve large coalitions.

Another observation that may have considerable generality is that markets of this kind have a propensity to unravel backwards in time, with dates of appointment becoming earlier and earlier. This phenomenon seems to occur in very different kinds of markets. (Indeed, the membership drives of American sororities are called "rush" precisely because they have unraveled in this way [see Mongell and Roth, 1991]. Among the markets presently experiencing this sort of unraveling are some of those for newly graduated lawyers, about which I hope to have more to say in the future.) This suggests that the kind of models considered here may be useful tools with which to explore a wide variety of entry-level labor markets and related matching processes. Centralized markets have been a productive place to begin this kind of study, because they make it relatively straightforward to determine the "rules of the game" in the kind of detail demanded by game-theoretic analysis. However, with the experience gained from studying such markets, it should be possible to study decentralized markets in related ways.⁴²

procedures. Sir James Fraser (pers. comm., 10 July 1987), writing of the Edinburgh PRAMS, notes that "... one of the principles behind the Scheme is that a student is committed to accept the Unit to which he is allotted. Rarely, this formal agreement is broken, more commonly by applicants from outside Edinburgh..."

⁴¹Of course, there may be formidable and perhaps intractable problems of coordination, having to do with different schedules and jurisdictions, involved in setting up such a national market.

⁴²A theoretical question related to the study of decentralized markets has recently been resolved in a paper by Roth and John Vande Vate (1990), which

In closing, a few words are probably in order about the methodology this paper shares with the related empirical studies of matching in Roth (1984a) and Mongell and Roth (1991). The chief analytical tools (stability and strategic equilibrium) are a mix drawn from what has traditionally been called cooperative and noncooperative game theory. Their use together here should help make clear why this is not always a useful distinction, since these two approaches to game theory are complements, rather than substitutes. While the rules of the game must be identified in great detail in order to speak about equilibrium, the stability of a matching may be discussed somewhat independently of the specific rules of the market. One of the phenomena that emerges from these studies is that, when the outcomes are unstable, agents have incentives to change the rules of the game, as when they decide to introduce a centralized matching procedure or to defect from one. In principle, all such decisions could be modeled in terms of some larger, all-encompassing game, and much of the contemporary theoretical literature in game theory takes this point of view. However, this will seldom be an option in modeling complex real systems, whose rules need to be determined by observation (and determining rules by observation will often be subject to uncertainty, since as has been seen, formal rules sometimes turn out to be less than fully binding, while informal unwritten rules may impose real constraints). If game theory is to play as important a role in empirical economics as it already plays in economic theory, increased attention must be paid to modeling complex games whose rules can only be observed imperfectly.

APPENDIX

Proposition 10 below is an impossibility result that strengthens the conclusions of

Proposition 4 and shows that no priority matching scheme always produces stable matchings. An arbitrary priority matching scheme is given by a priority function $f: \mathbf{N} \cup \{x\} \times \mathbf{N} \cup \{x\} \rightarrow \mathbf{N} \cup \{x\}$, where \mathbf{N} is the set of positive integers and x indicates that a match is unacceptable. Thus, if a_{ij} is student j 's rank in consultant i 's ordering (i.e., $a_{ij} = k$ means student j is consultant i 's k th choice) and b_{ji} is consultant i 's rank in student j 's ordering, $f(a_{ij}, x) = f(x, b_{ji}) = x$, and $f(a_{ij}, b_{ji}) \in \mathbf{N}$, where $f(a_{ij}, b_{ji})$ is the priority of the match between i and j . The matching μ_f is created by performing all first-priority matches [$f(a_{ij}, b_{ji}) = 1$], followed by all second-priority matches among those agents not yet matched, and so forth. Note that the function f restricted to acceptable matches must be one-to-one (i.e., $f: \mathbf{N} \times \mathbf{N} \rightarrow \mathbf{N}$ is one-to-one) because, for example, (1,2) matches cannot have the same priority as (2,1) matches, since a given consultant might be part of a (1,2) match with s_1 and a (2,1) match with s_2 . Without loss of generality (by taking a monotone transformation of f if necessary), one may also assume that the function f is onto \mathbf{N} [i.e., that there are values of a_{ij} and b_{ji} for which $f(a_{ij}, b_{ji})$ equals 1, 2, etc., with no gaps]. The impossibility result can now be stated.

PROPOSITION 10: *No priority matching function f always produces stable matches μ_f .*

PROOF:

It will be sufficient to prove the proposition for the symmetric case of one-to-one matching. First observe that if (1,1) matches are not first-priority (i.e., if $f(1,1) \neq 1$) then f does not always produce stable matchings. To see this, suppose (i,j) matches are first-priority, and suppose $j > 1$. Let $a_{21} = i$ and $b_{12} = j$, so C_2 and s_1 are matched at the matching μ_f ; but let $a_{11} = 1 = b_{11}$, so C_1 and s_1 are a (1,1) match. Then μ_f is unstable with respect to C_1 and s_1 , which proves the first observation.

Next observe that if $(i,j) \neq (1,1)$ is second-priority and $i > 1$ then μ_f may be unstable, even if (1,1) matches are first-priority. To see this, let (1,1) matches be first-priority and let C_1 , C_2 , s_1 , s_2 , and s_3 be

shows that a wide class of random processes converge with probability 1 to a stable matching. Some preliminary results on strategic behavior in such markets are contained in Roth and Vande Vate (1991).

such that $a_{11} = 1$, $a_{12} = i > 1$, $b_{11} = 2$, $b_{12} = 1$, $a_{23} = 1$, $b_{21} = j$, and $b_{32} = 1$. Then C_2 and s_3 are matched in the $(1, 1)$ phase, and C_1 , s_1 , and s_2 are not (if $j \neq 1$, let $b_{22} = 1$). Therefore, C_1 and s_2 are matched in the (i, j) phase, and the resulting matching is unstable with respect to C_1 and s_1 , which proves the second observation.

A symmetric argument shows that unstable matchings may result unless $j = 1$ also. However, since $(i, j) \neq (1, 1)$, this completes the proof of the proposition.

A formal model of matching sufficiently general for all the cases covered here is the following. There are two sets of agents; $C = \{C_1, \dots, C_n\}$ and $S = \{s_1, \dots, s_m\}$, namely consultants and students. Each consultant C seeks to fill no more than some number q_C of positions, and each student s seeks to obtain no more than $q_s = 2$ positions. A *matching* is a function from $C \cup S$ into the set of all subsets of $C \cup S$ such that, for every C in C and s in S :

- 1) $\mu(C)$ is contained in S and $\mu(s)$ is contained in C (either set may be empty);
- 2) $|\mu(C)| \leq q_C$ for all C in C ;
- 3) $|\mu(s)| \leq q_s = 2$ for all s in S ;
- 4) s is in $\mu(C)$ if and only if C is in $\mu(s)$.

Each agent A has preferences over sets of agents on the other side of the market that determine a choice function Ch_A as in Section II. Each agent's preferences over matchings are sensitive only to his own part of the matching, so that if μ and ν are matchings such that $\mu(A) = \nu(A)$, then agent A is indifferent between μ and ν .

An agent A has *responsive* preferences if he has a rank-ordering of the individual agents on the other side of the market such that, for any subset T of such agents, $Ch_A(T)$ is the set of the q_A highest-ranked acceptable individuals in T (or all the acceptable individuals in T if there are fewer than q_A). Note that this is a slightly weaker definition than that given in Section I, since this definition deals only with the agents' choice sets and ignores preferences between subsets T that would not arise as choices from a common set. Note also that responsive prefer-

ences are substitutable, as defined in Section II.

The following example proves the second part of Proposition 2, namely that stable matchings need not be in the core, nor even be Pareto optimal, in many-to-many matching with responsive preferences.

Example 3. There are four consultants, each of whom seeks two house officers, and four students, each of whom seeks two positions. The (responsive) preferences of the agents are given by the following rank-orderings of acceptable subsets:

- s_1 : $\{C_1, C_2\}$, $\{C_1, C_3\}$, $\{C_1, C_4\}$, $\{C_2, C_3\}$,
 $\{C_2, C_4\}$, $\{C_3, C_4\}$, $\{C_1\}$, $\{C_2\}$, $\{C_3\}$, $\{C_4\}$
 s_2 : $\{C_2, C_1\}$, $\{C_2, C_4\}$, $\{C_2, C_3\}$, $\{C_1, C_4\}$,
 $\{C_1, C_3\}$, $\{C_4, C_3\}$, $\{C_2\}$, $\{C_1\}$, $\{C_4\}$, $\{C_3\}$
 s_3 : $\{C_3, C_4\}$, $\{C_3, C_1\}$, $\{C_3, C_2\}$, $\{C_4, C_1\}$,
 $\{C_4, C_2\}$, $\{C_1, C_2\}$, $\{C_3\}$, $\{C_4\}$, $\{C_1\}$, $\{C_2\}$
 s_4 : $\{C_4, C_3\}$, $\{C_4, C_2\}$, $\{C_4, C_1\}$, $\{C_3, C_2\}$,
 $\{C_3, C_1\}$, $\{C_2, C_1\}$, $\{C_4\}$, $\{C_3\}$, $\{C_2\}$, $\{C_1\}$
 C_1 : $\{s_4, s_3\}$, $\{s_4, s_2\}$, $\{s_4, s_1\}$, $\{s_3, s_2\}$, $\{s_3, s_1\}$,
 $\{s_2, s_1\}$, $\{s_4\}$, $\{s_3\}$, $\{s_2\}$, $\{s_1\}$
 C_2 : $\{s_3, s_4\}$, $\{s_3, s_1\}$, $\{s_3, s_2\}$, $\{s_4, s_1\}$, $\{s_4, s_2\}$,
 $\{s_1, s_2\}$, $\{s_3\}$, $\{s_4\}$, $\{s_1\}$, $\{s_2\}$
 C_3 : $\{s_2, s_1\}$, $\{s_2, s_4\}$, $\{s_2, s_3\}$, $\{s_1, s_4\}$, $\{s_1, s_3\}$,
 $\{s_4, s_3\}$, $\{s_2\}$, $\{s_1\}$, $\{s_4\}$, $\{s_3\}$
 C_4 : $\{s_1, s_2\}$, $\{s_1, s_3\}$, $\{s_1, s_4\}$, $\{s_2, s_3\}$, $\{s_2, s_4\}$,
 $\{s_3, s_4\}$, $\{s_1\}$, $\{s_2\}$, $\{s_3\}$, $\{s_4\}$.

Each agent's preferences over individuals can be read from the last four (singleton) entries in his preference list P over sets. The matching μ that matches each agent to his fourth-choice set of agents [emphasized in boldface in the preferences; i.e., $\mu(s_1) = \{C_2, C_3\}$, $\mu(C_1) = \{s_3, s_2\}$, etc.] is stable. To see this, note that for each s_i , all improvements on $\mu(s_i)$ involve consultant C_i , but C_i is not interested in dealing with s_i alone, since s_i is C_i 's last-choice individual. However, μ is not in the core and is not Pareto optimal, since it is dominated by the matching μ' that gives each agent his third-choice set of agents.⁴³

⁴³In each case, the responsive preferences have been chosen so that an agent prefers to be matched with the set consisting of his first- and fourth-choice

To adapt the example to the requirement that students must obtain one medical and one surgical position, suppose C_1 and C_3 offer surgical positions, while C_2 and C_4 offer medical positions. Then, the student preferences given above can be modified by simply deleting from the preference lists the unacceptable sets $\{C_2, C_4\}$ and $\{C_1, C_3\}$, which leaves the conclusion of the example unchanged. Similarly, by deleting the set $\{s_1, s_2\}$ from the preferences of C_4 (who might be a urological surgeon from Edinburgh), we obtain the kind of substitutable preferences that result when a gender quota is invoked.

Proposition 3 (and therefore the first part of Proposition 2) is proved next.

PROPOSITION 3: *In the many-to-many matching model in which all agents have substitutable preferences, the set of stable matchings is nonempty.*

The proof will use the following algorithm.

Step 1: Each C makes offers to every s in the set $Q(C, 1) = \text{Ch}_C(S)$.

...

Step k : (i) Each s rejects any offers received so far that are not in the set $\text{Ch}_s(O(s, k-1))$, where $O(s, k-1)$ is the set of offers s has received in steps $1, \dots, k-1$ and not yet rejected by the end of step $k-1$. (ii) Each C who has received at least one rejection in part (i) of step k and who now has (nonrejected) offers outstanding to the students in the set $N(C, k)$, and has been rejected in steps $1, \dots, k$ by the students in the set $R(C, k)$, makes (or renews) offers to his most preferred set of students $Q(C, k)$ in the collection of sets $\{Q \subset S \setminus R(C, k) \text{ such that}$

$Q \text{ contains } N(C, k)\}$. (Subtraction of sets is denoted by \setminus .) That is, C 's outstanding offers at the end of step k include all those issued at previous steps and not yet rejected and none of those that have already been rejected. (iii) The algorithm stops at any step $k = T$ at which no rejections are issued, and the resulting matching is μ such that $\mu(C) = N(C, T)$ for each C in C .

The proof, which involves showing that the matching μ resulting from the algorithm is stable (as defined in Section II) when preferences are substitutable, makes use of the following lemmas. The first says that the requirement that consultants keep open all offers that have not been rejected does not constrain them: at each step, they issue offers to their choice set from among those that have not yet rejected them. (The lemmas, like the algorithm, are stated as if preferences are strict so that there is always a unique choice set. This assumption is for notational convenience only; if it were relaxed, it would be necessary to denote a choice set, meaning one among the collection of most-preferred sets, instead of the choice set as below.)

LEMMA 1: *For each step $k = 1, \dots, T$ and each C in C ,*

$$Q(C, k) = \text{Ch}_C(S \setminus R(C, k)).$$

PROOF:

For $k = 1$, $R(C, 1) = \emptyset$, so $Q(C, 1) = \text{Ch}_C(S \setminus R(C, 1))$. Suppose now, inductively, that $Q(C, j) = \text{Ch}_C(S \setminus R(C, j-1))$ for $j = 1, \dots, k-1$. Then, if s is in $N(C, k)$, s is in $Q(C, k-1) = \text{Ch}_C(S \setminus R(C, k-1))$; but $\{S \setminus R(C, k-1)\}$ contains $\{S \setminus R(C, k)\}$, and s is an element of $\{S \setminus R(C, k)\}$, so substitutability implies that s is an element of $\text{Ch}_C(S \setminus R(C, k))$ [i.e., that $\text{Ch}_C(S \setminus R(C, k))$ contains $N(C, k)$]. Therefore, $Q(C, k) = \text{Ch}_C(S \setminus R(C, k))$, which completes the proof by induction.

The next lemma says that, no matter what the sequence of offers a student receives, he

individuals rather than his second and third choices. This is not a consequence of responsiveness; it is easy to check that reversing this preference would still leave the preferences responsive.

never regrets having rejected earlier offers when he discovers what the later offers are.

LEMMA 2: *For each s in S , if the algorithm stops at step T ,*

$$\begin{aligned}\mu(s) &\equiv O(s, T-1) \\ &\equiv \text{Ch}_s(O(s, T-1)) \\ &= \text{Ch}_s\left(\bigcup_{k=2}^T O(s, k-1)\right).\end{aligned}$$

PROOF:

Suppose that C is an element of $\text{Ch}_s(\bigcup_k O(s, k-1))$, and let k be such that C is an element of $O(s, k-1)$. Then by substitutability, C is an element of $\text{Ch}_s(O(s, k-1))$. Therefore, s does not reject C at any step of the algorithm, so C is an element of $O(s, T-1)$, and $\text{Ch}_s(\bigcup_k O(s, k-1))$ is a subset of $\mu(s)$. However, $\mu(s) = \text{Ch}_s(O(s, T-1))$ cannot be strictly preferred to its subset $\text{Ch}_s(\bigcup_k O(s, k-1))$, and so when preferences are strict the two sets must be equal (and when preferences are not strict the choice sets can always be selected so that the two sets are equal).

The proof of Proposition 3 is now simple.

PROOF OF PROPOSITION 3:

The lemmas imply that $\mu(s) = \text{Ch}_s(\mu(s))$ for all students s , and $\mu(C) = \text{Ch}_C(\mu(C))$ for all consultants C . Now consider a student s and consultant C , not matched to one another at μ , such that s is an element of $\text{Ch}_C(\mu(C) \cup s)$. Lemma 1 implies that, during the course of the algorithm, C made an offer to s and was rejected. Then, Lemma 2 implies that $\text{Ch}_s(\mu(s) \cup C) = \mu(s)$. Therefore, μ is stable.

The above proof demonstrates that the set of stable matchings is nonempty in many-to-two matching with substitutable preferences. Proposition 6 states that preferences constrained to obey constraints 1 or 2 described in Section IV are substitutable,

while those constrained to obey 1 and 3 are not.

PROOF OF PROPOSITION 6:

First consider an agent A with quota q_A ($= 2$ if A is a student) whose preferences are responsive except insofar as they satisfy either constraint 1 (if A is a student) or 2 (in which case A must be a consultant surgeon). That is, A has a rank-ordering over the agents in the opposite set, which can be represented as a set of the form $B = B_1 \cup B_2$, and A finds it unacceptable to be matched to any subset S of B that contains more than one member of B_2 (and, if A is a student, more than one member of B_1). Thus, for any subset T of B , $\text{Ch}_A(T)$ is a set consisting of the q_A highest-rank acceptable members of T (or all of them if there are fewer than q_A), providing that these contain no more than one member of B_2 (or B_1), and otherwise it is a set consisting of the highest-ranked member of $B_2 \cap T$ and the $q_A - 1$ highest-ranked acceptable members of $B_1 \cap T$ (or all of them if there are fewer than $q_A - 1$). Now it is easy to check that, for any t in B_1 or B_2 and $t' \neq t$ in B_1 or B_2 , if t is in $\text{Ch}_A(T)$ then it is also in $\text{Ch}_A(T - t')$. This completes the proof of the proposition for the case of preferences satisfying constraints 1 or 2.

Now consider a student s whose preferences satisfy constraints 1 and 3 but are otherwise responsive to a rank-ordering over individual consultants, and let this rank-ordering be C_1, C_2, C_3, C_4 , where C_1 and C_3 are teaching-hospital positions in medicine and surgery, respectively, and where C_2 and C_4 are nonteaching hospital positions in medicine and surgery, respectively. Then, it is possible to have $\text{Ch}_s(\{C_1, C_2, C_3, C_4\}) = \{C_1, C_4\}$ but $\text{Ch}_s(\{C_2, C_3, C_4\}) = \{C_2, C_3\}$, which does not contain C_4 , so these preferences are not substitutable. This completes the proof of Proposition 6.

PROOF OF PROPOSITION 7:

Recall that each consultant in the deferred-acceptance algorithm as modified at Edinburgh submits a rank-ordering of individual students. Following a rejection in the

algorithm, a consultant extends an offer to the highest-ranked (eligible) student who has not yet received an offer from him, where only male students may be eligible if the consultant is a surgeon who has exercised his option to limit the number of female house officers. Each student has submitted separate lists of medical and surgical positions and rejects all but the best offer in each category.

It is immediate from the above description of the algorithm and from Proposition 6 that this conforms to the version of the algorithm used here to prove Proposition 3, if the consultants have responsive preferences that are lexicographic in their lower-ranked (more preferred) candidates (e.g., a consultant C whose rank-ordering is $s_1, s_2, s_3, s_4, \dots$ must prefer $\{s_1, s_4\}$ to $\{s_2, s_3\}$ in order for the algorithm to correspond to that used to prove Proposition 3). Thus, if the consultants have such preferences, the matching is stable. However, suppose they have responsive preferences that are not lexicographic in this way (e.g., suppose the above consultant C prefers $\{s_2, s_3\}$ to $\{s_1, s_4\}$). If there were a student s such that the resulting matching was unstable with respect to (s, C) , then the matching would be unstable with respect to *any* preferences that are responsive to C 's rank-ordering of individual students. Since this cannot be, the resulting matching is stable.

The following example shows that the set of stable matchings may be empty when student preferences obey both constraints 1 and 3 from Section IV:

Example 4. Consider a market that includes four students applying to consultants MT_1 , ST_1 , ST_2 , SNT , MNT , where M stands for a medical position, S for a surgical position, T for a teaching-hospital position and NT for a nonteaching-hospital position. For simplicity, each consultant seeks only one student. Each student seeks one medical and one surgical position and may not hold more than one teaching-hospital position. The preferences over individual students and positions are given by the following rank

orderings:

s_1 : SNT, ST_2, MT_1	MT_1 : s_1, s_2
s_2 : MT_1, ST_1	ST_1 : s_2, s_3
s_3 : ST_1, SNT	ST_2 : s_1, s_4
s_4 : ST_2	SNT : s_3, s_1, s_2
	MNT : s_1, s_2

To see that every matching μ is unstable in this example, first suppose that μ is stable and that $\mu(SNT) \neq s_1$. Stability implies $\mu(SNT) = s_3$, which implies $\mu(ST_1) = s_2$, which implies $\mu(MT_1) = s_1$, which implies that μ is unstable with respect to (s_1, ST_2) , a contradiction. Thus, if μ is stable, $\mu(SNT) = s_1$. However, this implies $\mu(ST_1) = s_3$, which implies $\mu(MT_1) = s_2$, which implies that μ is unstable with respect to (s_1, MT_1) , which is the contradiction needed to prove that no matching is stable.

REFERENCES

- Acheson, E. D., "Preregistration House-Officers" (letter), *The Lancet*, 14 April 1984 (Vol. I, No. 8381), 852.
- Alexander-Williams, J. and Stephenson, Ivor G., "Appointment of Preregistration House Officers," *British Medical Journal*, 9 June 1973 (Vol. 2, No. 5866), 605-6.
- Blair, Charles, "The Lattice Structure of the Set of Stable Matchings with Multiple Partners," *Mathematics of Operations Research*, November 1988, 13, 619-28.
- Clayden, A. D. and Parkhouse, James, "Allocation of Preregistration Posts," *British Journal of Medical Education*, March 1971, 5, 5-12.
- Doig, A. and Munday, G., "A Coordinated Scheme for the Allocation of Preregistration House-Officer Posts," *The Lancet*, 21 June 1969 (Vol. I, No. 7608), 1250-2.
- Gale, David and Shapley, Lloyd, "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, January 1962, 69, 9-15.
- Gillard, J. H. and Dent, T. H. S., "The Allocation of House Officer Posts: A UK Sur-

- vey," *Medical Education*, July 1988, 22, 342-4.
- Kelso, Alexander J., Jr. and Crawford, Vincent P., "Job Matching, Coalition Formation, and Gross Substitutes," *Econometrica*, November 1982, 50, 1483-1504.
- Leishman, A. G. and Ryan, R. P., "Appointment of Provisionally Registered House-Officers by Computer Match," *The Lancet*, 29 August 1970 (Vol. I, No. 7670), 459-61.
- McVitie, D. G. and Wilson, L. B., (1970a) "Stable Marriage Assignments for Unequal Sets," *BIT*, 1970, 10 (3), 295-309.
- _____, and _____, (1970b) "The Application of the Stable Marriage Assignment to University Admissions," *Operational Research Quarterly*, December 1970, 21, 425-33.
- _____, and _____, "The Stable Marriage Problem," *Communications of the ACM*, July 1971, 14, 486-92.
- Mongell, Susan and Roth, Alvin E., "Sorority Rush as a Two-Sided Matching Mechanism," *American Economic Review*, June 1991, 81, 441-64.
- Roth, Alvin E., "The Economics of Matching: Stability and Incentives," *Mathematics of Operations Research*, November 1982, 7, 617-28.
- _____, (1984a) "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory," *Journal of Political Economy*, December 1984, 92, 991-1016.
- _____, (1984b) "Misrepresentation and Stability in the Marriage Problem," *Journal of Economic Theory*, December 1984, 34, 383-7.
- _____, (1985a) "The College Admissions Problem Is Not Equivalent to the Marriage Problem," *Journal of Economic Theory*, August 1985, 36, 277-88.
- _____, (1985b) "Common and Conflicting Interests in Two-Sided Matching Markets," *European Economic Review* (special issue: Market Competition, Conflict, and Collusion), February 1985, 27, 75-96.
- _____, "Two Sided Matching with Incomplete Information about Others' Preferences," *Games and Economic Behavior*, June 1989, 1, 191-209.
- _____, "New Physicians in the U.S. and the U.K.: A Natural Experiment in the Organization of Labor Markets," *Science*, 14 December 1990, 250, 1524-8.
- _____, and Sotomayor, Marilda, *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Econometric Society Monograph Series, Cambridge: Cambridge University Press, 1990.
- _____, and Vande Vate, John H., "Random Paths to Stability in Two-Sided Matching," *Econometrica*, November 1990, 58, 1475-80.
- _____, and _____, "Incentives in Two-Sided Matching with Random Stable Mechanisms," *Economic Theory*, January 1991, 1, 31-44.
- Shah, A. R. and Farrow, S. C., "Pre-registration House Appointments: A Computer Aided Allocation Scheme," *Medical Education*, November 1976, 10, 474-9.
- Townsend, H. R. A., "Pre-registration House Appointments: A Computer-Aided Allocation Scheme" (letter), *Medical Education*, March 1977, 11, 160-1.
- _____, *PRAMS.80 Manual*, Cirencester, U.K.: Special Medical Micro Software Ltd., 1981.
- Yellowlees, Henry, "Difficulties in Obtaining Preregistration Posts" (letter), *The Lancet*, 26 November 1983 (Vol. II, No. 8361), 1254.
- Department of Health and Social Security (U.K.), *Report of the Pre-registration House Officers Working Group*, unpublished memorandum, 1987.

Sorority Rush as a Two-Sided Matching Mechanism.

By SUSAN MONGELL AND ALVIN E. ROTH*

The history and organization of the membership recruitment process of American sororities is studied. Like entry-level labor markets studied previously, this process experienced failures that led to the adoption of a centralized matching procedure in which a matching is determined on the basis of preference lists submitted by the agents. Analysis of the rules of the match and of preference lists from 21 matches reveals an unstable matching procedure that gives agents incentives to behave strategically. The analysis also shows how the agents act on these incentives and how the resulting strategic behavior has contributed to the longevity of the matching system and to the stability of the resulting matches. (JEL C78, D00, J41)

This paper concerns the formal process by which women at American universities join the social organizations called sororities. The history of this process, of the problems it has encountered, and how it has evolved to meet them, have striking similarities to (as well as important differences from) the history and organization of the American labor market for medical interns (see Roth, 1984a) and the several similar entry-level labor markets for physicians in the United Kingdom (see Roth, 1991). Thus, by studying this process, we hope to learn more about other matching processes and hope to assess the generality of various hypotheses about them.

In the medical labor markets, competition for newly graduating medical students and for desirable positions caused the dates at which appointments were finalized to unravel in time, so that by the 1940's in the United States and by the 1960's in the United Kingdom, postgraduation employ-

ment was often arranged well over a year (and sometimes over two years) in advance of graduation. Similarly, by the latter part of the last century, entry into fraternities and sororities, initially reserved for college seniors, had worked its way backward to the freshman class, and in some cases membership was arranged well before matriculation. (This aspect of the competition for members appears to be the origin of the term "rushing," as these membership drives are now called.)

Largely in response to the problems arising out of this kind of unraveling, the parties involved in the different medical labor markets eventually agreed to try a variety of centralized matching procedures, in which participants would not sort themselves out individually but would instead submit rank-orderings of their choices to a central clearinghouse, which would use this information to match students to jobs. Similarly, about 60 years ago, the umbrella organization of American sororities recommended that a centralized procedure be used to match students to sororities on college campuses.

The basic mechanism used to process the rank-orderings submitted by students and sororities is called the "preferential bidding system" (PBS), and it remains in use today. This paper analyzes the PBS algorithm, the setting in which it is employed, the incentives it gives to students and sororities, and the matchings that result.

*Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260. Because of the requirement that the campuses should remain anonymous, we are unable to thank by name the many administrators without whose help this study could not have proceeded. We have received helpful comments from Patty Beeson. This work has been supported by grants from the Alfred P. Sloan Foundation, the National Science Foundation, and the Office of Naval Research.

This analysis reveals that the PBS algorithm is different in an important way from the algorithm around which the American medical market is organized and the algorithms around which some of the most successful and long-lived of the medical markets in the United Kingdom are organized. Those other algorithms have the property that they produce matchings that are *stable*, in the sense that, for markets of the kind considered here, they are in the core of the market, when agents state their true preferences.¹ The PBS algorithm does not have this property: the matchings it produces are stable in the preliminary market in which the algorithm operates, but they are not in the core of the market as a whole. Furthermore, for many configurations of preferences, the algorithm fails to produce a matching at all.

Nevertheless, when we examine data from 21 rushes on four campuses,² we observe only one such failure. A large part of the explanation appears after further examination of the data, which makes clear that the submitted preferences are unlikely to correspond to the true full preferences of the students. Instead, the observed pattern of preferences corresponds to what we would expect to see if the students respond strategically to the incentives induced by the matching procedure. When students do respond this way, the PBS procedure will not fail, and the resulting matching will be stable.

I. A Brief History

The first Greek-letter sorority was founded in 1870 (James Brown, 1920). A sorority may be present on campuses throughout the United States, and each

sorority location is called a chapter. In the literature of fraternities and sororities, from which we will sometimes quote, "fraternity" is used to mean either the all-male or the all-female social organizations, while "sorority" refers to the all-female organizations. Many sororities have joined a national organization, the National Panhellenic Conference (NPC), which consisted of 26 member sororities as of 1985. On each campus, all NPC sorority chapters are members of a College Panhellenic Council, the local governing body that determines rushing regulations.

Brown (1920 p. 14) described the early competition for members:

In the early days of the fraternities only seniors were admitted to membership, but the sharp rivalry for desirable men soon pushed the contest into the junior class, and so on down, until at some colleges it scarcely stops at the academy. The general rule is, however, that members shall be drawn from the four undergraduate classes. ... As the colleges usually open about the middle of September, the campaign for freshmen is then commenced and lasts until Christmas, when each chapter has secured its most desirable candidates. Where there is great rivalry, however, initiations take place all year round.

Earlier appointment dates were not the only evidence of competition:

Membership in two fraternities has been a source of trouble and vexation. It is almost universally forbidden. When it occurs between two chapters of different fraternities located at the same college, and a student leaves one and joins the other, it is termed "lifting," and such disloyalty is usually followed by expulsion. ... All of the fraternities now forbid this, although many years ago it was not uncommon. [Brown, 1920 pp. 15-16]

An early attempt to resolve these problems occurred in 1891, when the first meeting of sororities, in what was then called the

¹These markets involve many-to-one matching, since each student joins at most one sorority. In markets of many-to-many matching, stable matchings need not be in the core (see Roth, 1991).

²The data available to us come from campuses in which, loosely speaking, there is a "buyers' market" for sorority positions. We will discuss differences expected in "seller's markets."

Inter-Sorority Conference, was called to discuss intersorority cooperation. Although resolutions were passed decrying the practice of "lifting" and calling for the "abolition of the practice of pledging and initiating preparatory students" (National Panhellenic Council, 1983 p. 5), this had little effect. Similar sentiments were expressed in subsequent years, to equally little effect, and by 1928 the NPC was ready to turn to a centralized system of matching, and the first mention of the preferential bidding system appears.³ Francis Shephardson (1930 p. 8) reviews the events leading up to this:

The constant rivalry among chapters and the multiplication of fraternities have led in many cases to an indiscriminate scramble for members at the beginning of each year. Both fraternities and the colleges have perceived the danger of this sort of "rushing," as the contest for members is called, and are giving the subject thoughtful consideration. The deferred pledging of students until a fixed date and the deferred initiation of pledged members until they have completed a prescribed portion of their college course or secured a predetermined grade are both becoming common. Such procedure is in striking contrast with earlier custom in some of the larger Western and Southern colleges where, the preparatory schools being intimately connected with the colleges, "preps" were not only pledged, but initiated before they entered the college proper, or with the reprehensible custom which prevails in some places, where pledge pins are given out to boys in the high school or even in the grammar grades.

The preferential bidding system has since been incorporated into the recruiting activities of sororities, as described next.

II. The Organization of Recruiting Activities

The activities of a sorority seeking new members are called rush.⁴ There are two types of rush: formal rush and continuous open bidding. The NPC recommends "one formal rush period per year, held in the early fall, as close as possible to the start of the academic year, and conducted in as short a period of time as possible" (National Panhellenic Conference, 1979 p. 29).

Women participating in formal rush ("rushees") attend a sequence of parties designed to enable rushees and sororities to "narrow their choices gradually" (National Panhellenic Conference, 1979 p. 35). The first parties are "open houses" in which all sororities issue invitations to all rushees. In subsequent rounds, sororities issue invitations selectively. "Panhellenic strongly urges each sorority to re-invite ... only those rushees they are seriously considering for membership. This will enable both the rushee and the sororities to know 'how they stand' early in the formal rush period" (National Panhellenic Conference, 1979 p. 46). In each round, the number of sororities a rushee can attend is reduced. A rushee who receives more invitations than the number of parties permitted in a given round must decline, or "regret," the excess invitations. In the last round of invitational parties, the "preference parties," a rushee is usually permitted to attend only two or three parties. "Panhellenic strongly urges each sorority to invite *only* those rushees to the preference party to whom they will definitely issue a bid" (National Panhellenic Conference, 1979 p. 46).

After the last preference party, rushees indicate their sorority preferences on a card, which they sign. (A rushee who lists only a single sorority is said to have *suicided*.) Sororities similarly submit a preference or-

³See *National Panhellenic Review* (1985 p. 5) for a dated list of motions passed.

⁴The process described next is the recommended procedure appearing in the *"How To" Manual for College Panhellenics* (National Panhellenic Conference, 1979). While these rush procedures are not required, the essential features have been incorporated in each of the campuses we contacted.

dering of rushees. Once all preferences have been submitted, the PBS algorithm matches rushees to sororities.

Each sorority is eligible to be matched to up to *quota* (q) rushees during formal rush, where *quota* is "the number of rushees accepting at least one invitation to the first round of invitational parties, divided by the number of participating fraternities"⁵ (National Panhellenic Conference, 1979 p. 37).

Following the completion of the PBS algorithm, there is one more step in the formal rush process, which officially exists in two slightly different forms (and which in practice seems to vary somewhat more from campus to campus). Under the "quota-only" procedure, any sorority that has been assigned some number p of rushees by the PBS algorithm with $p < q$ is allowed to extend one additional set of at most $q - p$ bids to unmatched rushees. Under the "quota-plus" procedure, any sorority that has not been assigned q new members under the PBS algorithm or whose total membership $m + p$ (including the p new members) is below the total allowable chapter size, T (which is the same for all sororities on a given campus), is allowed to extend one additional set of at most $\max\{q - p, T - (m + p)\}$ bids to unmatched rushees. Rushees who were unmatched by the PBS algorithm are free to accept at most one of the bids they receive or to decline all such bids.

The results are announced on "pledge day," marking the end of formal rush. A rushee who enters formal rush by signing a preference card but who subsequently declines to join a sorority to which she has been matched, is not permitted to join another sorority for one year.

Continuous open bidding begins immediately after the close of formal rush. During continuous open bidding, any sorority that

has not received q new members or that has received q new members but is nevertheless below the total allowable chapter size is allowed to recruit additional members by simply extending them invitations to join. At this stage, sororities are not restricted to make a single set of bids but may recruit continuously until their membership reaches T (or, in the case of sororities whose initial membership m was greater than $T - q$, until they have recruited q new members).

This recruitment and matching process resembles those of the centralized medical labor markets (Roth, 1984a, 1991) mentioned in the Introduction: an information gathering period is followed by a centralized matching algorithm, which is followed by a decentralized "after-market." In the case of the medical labor markets, analysis of the matching algorithms is critical to understanding the matching process as a whole. We turn now to a detailed description of the PBS algorithm.

Rushees submit a "preference card" listing the sororities they would be willing to join, in order of preference. Sororities submit a "bid list" of rushees whom they would be willing to have as members. While a rushee can join no more than one sorority, every sorority is able to extend at least q (quota) invitations for new members through the formal rush process. Beyond the first-quota names, sororities list rushees in order of preference. These preference lists are used by the PBS algorithm to assign rushees to sororities. The instructions in Table 1 are from the manual "*How To*" for *College Panhellenics* (National Panhellenic Conference, 1979 pp. 41–2). These instructions are incomplete and contain ambiguous phrases, such as "This process is repeated as long as there is any possibility of a rushee receiving a bid from the fraternity of her first choice" and "When it becomes apparent that a rushee will not receive a bid from the fraternity of her first choice..." The NPC does have a pamphlet explaining the instructions of the PBS algorithm via an example to be conducted in a workshop. This example still does not handle some of the contingencies that may arise during an actual PBS execution.

⁵If this number is not an integer, it is rounded either up or down at the discretion of the individual supervising the rush. Quota can be rounded down without leaving some students unmatched, since rushees sometimes drop out of the formal rush process after the first round of invitational parties.

TABLE 1—SORORITY RUSHING INSTRUCTIONS:
THE PREFERENTIAL-BIDDING-SYSTEM ALGORITHM

Bid Lists:

1. At a specified time, each fraternity files with the Panhellenic Executive a list of women it wishes to bid.
 - a. Lists are in duplicate; one copy is used in bid matching, the other is returned to the chapter when the bid matching is completed.
 - b. The fraternity bid list should be on paper ruled into three columns:
 - Left-hand column—List in alphabetical order of fraternity's first choices up to the limit of quota.
 - Right-hand column—List in order of preference the fraternity's additional choices which may number as many as the chapter wishes to submit.
 - Center column—Is left blank, as this is the column in which the matched bids are entered.

As a bid is matched, the rushee's name is crossed off every fraternity's first or second list. Her name is entered in the center column of the fraternity list of the group to which she is being pledged.
2. Along with its bid lists, each fraternity brings to Panhellenic enough formal bids (in envelopes) for each woman to be pledged. These formal bids are to be addressed after bid matching is completed.

Procedure for Matching Bids:

1. Persons matching bids include the reader, the tabulator, and one alumna handling the bid list from her fraternity. Undergraduates are not to participate in bid matching.
2. Before bid matching begins, names of all rushees who chose not to sign a preference card should be crossed off all preference lists, and those lists adjusted to fill the space of these women.
3. Mechanics:
 - a. After alphabetizing the preference cards, the reader calls the rushee's name and her first choice. If the fraternity of her first choice has given her a bid on its first bid list, it is a matched bid, and all others should cross her from their list. If the rushee's name is not on the fraternity's first bid list, her preference card is temporarily laid aside. Names of rushees who list only one preference and are unmatched at the end of the first reading should be crossed off all other bid lists and their cards laid aside.
 - b. Each time a name is crossed off a fraternity's first bid list, if openings in the fraternity's pledge quota remain, a name from the fraternity's second bid list is added, in the listed order, to the bottom of the unmatched names remaining on the first list. *The number of unmatched names on the adjusted first bid list and the number of those pledged must always equal quota* (unless a chapter has run out of names to add from its second bid list).
 - c. The cards laid aside in step "a" are read again according to the first choice of the rushee. This process is repeated as long as there is any possibility of the rushee receiving a bid from the fraternity of her first choice.
 - d. Those cards remaining are those of rushees whose names are on the second bid list of the fraternities of their first choice.
 - e. When it becomes apparent a rushee will not receive a bid from the fraternity of her first choice, a rushee's second choice is then matched, if possible, in the above manner.
 - f. Any remaining cards are then read according to the rushee's third choice and the same procedure followed.
 - g. The tabulator reads the results, and all bid lists are reviewed for accuracy.
 - h. Unmatched bids—If a rushee's preference card has failed to match for a bid, the Panhellenic Executive may contact the rushee and ask if she will accept a bid from a fraternity not previously listed among her choices, if this other fraternity has her name on one of their bid lists. Any rushee not bid by any of her preference choices is eligible at any future time for rushing and pledging by any fraternity.

Unfilled Quotas—If a fraternity has failed to fill its quota through this bid matching in formal rush, it may be contacted by the Panhellenic Executive to ask if the fraternity wishes to extend a bid to anyone not originally on its bid lists.

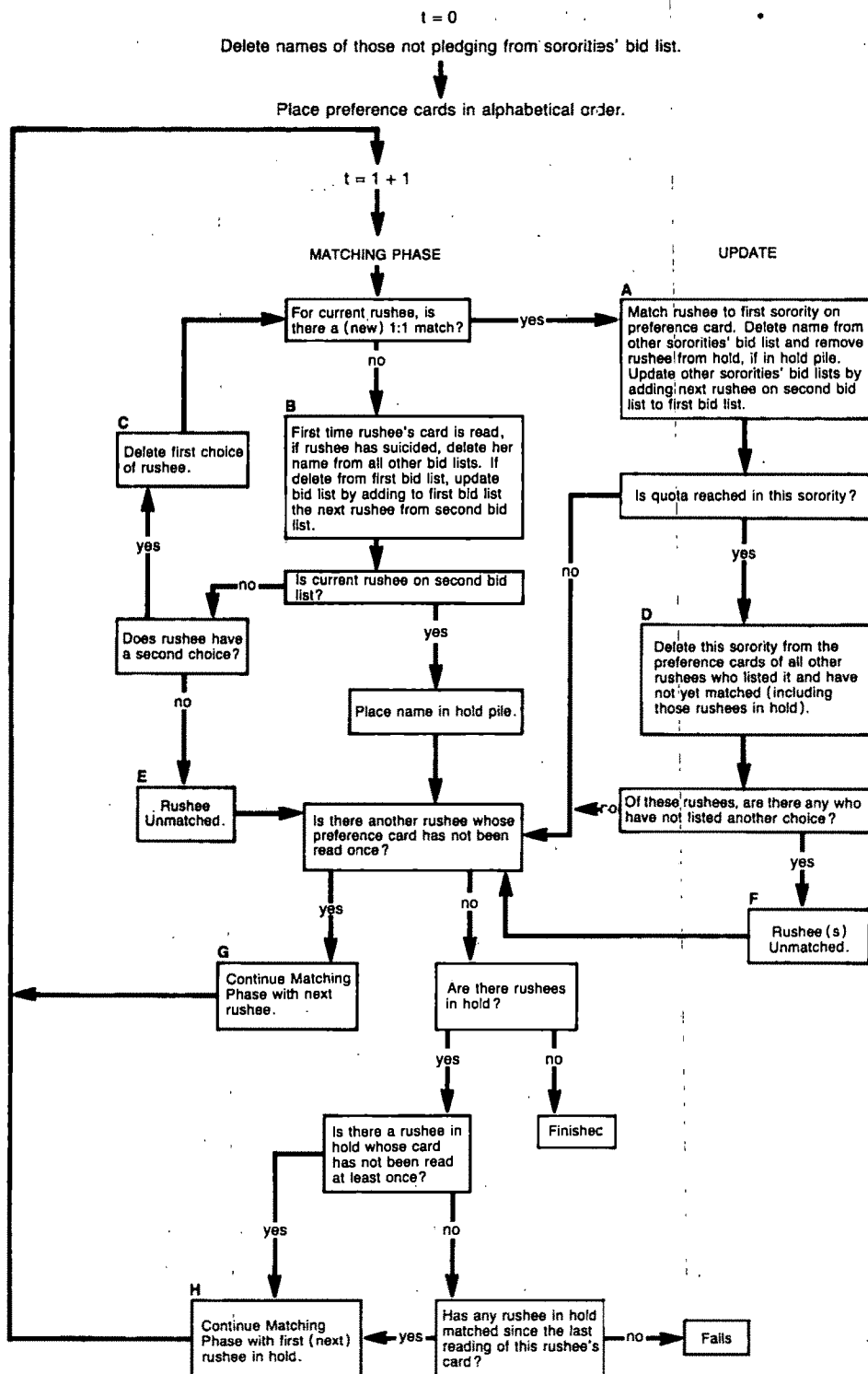


FIGURE 1. PBS-ALGORITHM FLOW CHART

When the instructions given for the PBS algorithm do not indicate what should be done with those rushees whose cards have been "laid aside," we will say that the algorithm "fails."⁶ When we examine the data, it will be seen that, in practice, the PBS algorithm very seldom fails. Indeed, the individuals in charge of administering the algorithm on each of the campuses from which our data are drawn were all initially unaware of the possibility of this kind of failure.⁷

Figure 1 is a flow chart of the PBS algorithm. No such flow chart is found in the sorority literature: this was compiled from both the literature mentioned above and interviews with individuals charged with supervising the matching process on some of the campuses contacted. The original bid list (before any rushees who have not signed a preference card have been deleted) is employed at step $t = 0$. Each time a rushee's preference card is read, t increases by 1.

The final step of the formal rush procedure, during which one set of additional bids may be made (see item h in Table 1 for one variation) has been omitted from the flowchart. Unlike the operation of the algorithm, such additional bids need input from the participants in addition to their initial preference lists. We will consider any such additional bids when we analyze the aftermath of the algorithm.

III. A Formal Model

The first elements of our formal model are two finite and disjoint sets of sororities

⁶For example, consider the case of two rushees and two sororities with $q = 1$. If rushee r_1 ranks sorority S_2 before S_1 , rushee r_2 ranks S_1 before S_2 , sorority S_1 ranks r_1 before r_2 , and S_2 ranks r_2 before r_1 , then both rushees will remain in hold, and the algorithm will fail.

⁷When subsequently presented with examples contrived so as to cause the algorithm to fail, these individuals suggested a variety of ad hoc procedures for restarting the algorithm and completing the matching procedure. Therefore, in the flow chart (Fig. 1), the box labelled "fails" can be viewed as a point in the algorithm in which the implementation would be different on different campuses.

and of rushees, $S = \{S_1, \dots, S_n\}$ and $R = \{r_1, \dots, r_m\}$, respectively. Each rushee has preferences over the sororities, and each sorority has preferences over the rushees. We will assume that these preferences are complete and transitive, with $P(S) = r_1, r_2, S, r_3, \dots$ denoting that sorority S prefers to enroll r_1 rather than r_2 , that it prefers to enroll either one of them rather than leave a position unfilled, and that all other rushees are unacceptable, in the sense that S prefers to leave a position unfilled rather than filling it with, say, rushee r_3 . Similarly, $P(r) = S_2, S_1, S_3, r, \dots$ represents the preferences of rushee r , indicating for example that the only positions the rushee would accept are those offered by S_2, S_1 , and S_3 , in that order. Sorority S is *acceptable* to rushee r if r prefers to be matched to S rather than to remain unmatched, and rushee r is *acceptable* to sorority S if S prefers to have r as a member rather than to leave a position unfilled.

The number of positions each sorority can fill during formal rush is q (quota). An outcome of the PBS algorithm is a matching of rushees to sororities, such that each rushee is matched to at most one sorority and each sorority is matched to at most q rushees. After formal rush (i.e., during the continuous open bidding which follows), each sorority S_k may have a different quota q_k , depending on its membership prior to the start of formal rush and on how many new members it has been matched to during formal rush.

A rushee who is not matched to any sorority will be modeled as "matched to herself," and a sorority with some number of unfilled positions will be matched to itself in each of those positions. A rushee is matched to a given sorority if and only if the sorority is matched to that rushee. To give a formal definition, first define, for any set X , an *unordered family of elements* of X to be a collection of elements, not necessarily distinct. Thus, an element of X may appear more than once, which distinguishes an unordered family from a subset of X .

Definition: A matching μ is a function from the set $S \cup R$ into the set of unordered

families of elements of $S \cup R$ such that:

- 1) $|\mu(r)| = 1$ for every rushee r and $\mu(r) = r$ if $\mu(r) \notin S$;
- 2) $|\mu(S)| = q$ for every sorority S , and if the number of rushees in $\mu(S)$, say p , is less than q , then $\mu(S)$ contains $q - p$ copies of S ;
- 3) $\mu(r) = S$ if and only if r is in $\mu(S)$.

Thus, $\mu(r_1) = S$ denotes that rushee r_1 is enrolled at sorority S at the matching μ , and $\mu(S) = \{r_1, r_3, S, S\}$ denotes that sorority S , with $q = 4$, enrolls rushees r_1 and r_3 and has two positions unfilled. (When sororities have different quotas, q_k replaces q for each sorority S_k .)

Each rushee's preferences over alternative matchings correspond exactly to her preferences over her own assignments at the two matchings. While we have described sororities' preferences over rushees, when q is greater than 1, each sorority must be able to compare groups of rushees in order to compare alternative matchings, and we have yet to describe the preferences of sororities over groups of rushees. (Until we have described sororities' preferences over matchings, our model will not be a well-defined game.)

The simplest assumption connecting sororities' preferences over groups of rushees to their preferences over individual rushees is one insuring that, for example, if $\mu(S)$ assigns sorority S its third- and fourth-choice rushees, and $\mu'(S)$ assigns it its second- and fourth-choice rushees, then sorority S prefers $\mu'(S)$ to $\mu(S)$. Specifically, let $P^\#(S)$ denote the preference relation of sorority S over all assignments $\mu(S)$ it could receive at some matching μ . A sorority S 's preferences $P^\#(S)$ will be called "responsive" to its preferences $P(S)$ over individual rushees if, for any two assignments that differ in only one rushee, it prefers the assignment containing the more preferred rushee, as described formally in the following:

Definition: The preference relation $P^\#(S)$ over groups of rushees is responsive [to the

preferences $P(S)$ over individual rushees] if, whenever $\mu'(S) = \mu(S) \cup \{r_k\} \setminus \{\sigma\}$ for σ in $\mu(S)$ and r_k not in $\mu(S)$, then S prefers $\mu'(S)$ to $\mu(S)$ [under $P^\#(S)$] if and only if S prefers r_k to σ [under $P(S)$]. (Subtraction of sets is denoted by \setminus .)

Note that S may be indifferent between distinct assignments $\mu(S)$ and $\mu'(S)$ even if S has strict preferences over individual rushees.

A matching μ is *individually irrational* if $\mu(r) = S$ for some rushee r and sorority S with either r unacceptable to S or S unacceptable to r . Such a matching is *blocked* by the unhappy agent. This terminology reflects that the rules allow every agent to withhold her (or its) consent from such a match. Similarly, a sorority S and rushee r will be said to *block* a matching μ together if they are not matched to one another at μ but would both prefer to be matched to one another rather than to (one of) their present assignments. That is, μ is blocked by the sorority-rushee pair (S, r) if $\mu(r) \neq S$ and if r prefers S to $\mu(r)$ and S prefers r to σ for some σ in $\mu(S)$. (Note that σ may equal either some rushee r' in $\mu(S)$ or, if at least one of sorority S 's positions is unfilled at $\mu(S)$, σ may equal S .) Matchings blocked by an individual or by a pair of agents are unstable in the sense that there are agents with the incentive and the power to disrupt such matchings.

Definition: A matching μ is stable if it is not blocked by any individual agent or any sorority-rushee pair.

It is not obvious that this definition will be adequate, since we might need to consider coalitions consisting of sororities and several rushees (all of whom might be able to enroll at the sorority). However, it can be shown that considering larger coalitions would not change the set of stable outcomes, which equals the core of the game with respect to weak domination (see Roth, 1985b; Roth and Marilda Sotomayor, 1990).

Sets S of sororities and R of rushees, together with a vector P of preferences, one

for each agent, constitute a *matching market*.⁸ In what follows, we will assume for simplicity that all preferences over individuals are strict (i.e., that no sorority is indifferent between two acceptable rushees and no rushee is indifferent between two acceptable sororities). (We do not assume that sororities may not be indifferent between different groups of rushees.)

IV. Analysis of the Algorithm

We begin with a model of the market up to the conclusion of the PBS algorithm: in this part of the market, each sorority may admit q new members (q is the same for all sororities). We will refer to this as the market with quota q .

Some notation will help describe the working of the algorithm. Denote by $x'(r_i) = S_j$ that rushee r_i was matched to sorority S_j during step t , where a *step* is the working of the algorithm associated with a reading of a single rushee's preference card. Denote by $x'(r_i) = r_i$ that rushee r_i was assigned as unmatched during step t . At step t when a rushee r_i 's preference card is read, if r_i neither matches to a sorority nor is assigned as unmatched, then her preference card will be placed in "hold" and will be reread after all other rushees who have yet to be assigned (as unmatched or matched to a sorority) have had their preference cards read. Finally, denote by $x(r) = S$ that rushee r was matched to sorority S at some step of the algorithm, and similarly denote by $x(r) = r$ that rushee r was assigned to be unmatched; define $x(S)$ to be the set of all rushees assigned to S (i.e., $x(S) = \{r | x(r) = S\}$). Note that x is not a matching, because it is not defined for all rushees, but only for

those not left in hold when the algorithm ends, and because $|x(S)|$ may be less than q (the remaining positions are not filled with copies of S).

Sororities indicate preferences by listing rushees on a first bid list of no more than q names and a second bid list. Denote by $r_i \in Q_t(S_k)$ that rushee r_i is listed on the first bid list of sorority S_k at step t in the algorithm. That rushee r_j is listed on the second bid list of S_k at step t in the algorithm is denoted by $r_j \in Q_t^+(S_k)$. For each sorority, the bid list at step $t=0$ is the original bid list.

Consider next what the algorithm does when confronted with preferences for which it is well-defined (i.e., for which it does not fail to produce a matching).

THEOREM 1: *If no rushees are left in "hold" at the end of the PBS algorithm, its outcome is stable in the market with quota q .*

Theorem 1 is proved in the Appendix.

Furthermore, the PBS algorithm has the property that all its assignments are *inevitable*, in the sense that all rushees who match to sororities by the PBS algorithm must match to the same sorority at every stable outcome and rushees assigned as unmatched by the algorithm must be unmatched at every stable outcome in the market with quota q . That is, we have the following result (proved in the Appendix).

THEOREM 2: *The preferential-bidding-system algorithm only makes inevitable assignments in the market with quota q .*

An immediate consequence of the theorem is the following.

COROLLARY: *The preferential bidding system assigns all rushees only when there exists a unique stable outcome in the market with quota q .*⁹

⁸This matching market is an example of what is sometimes called a "college admissions" model (see David Gale and Lloyd Shapley, 1962). When quotas all equal 1, the model is symmetric between both sides of the market and is called the "marriage model." For many years, it was thought that the college admissions model was essentially equivalent to the marriage model. That this is not the case was shown in Roth (1985a). The model presented here is the college admissions model as reformulated in Roth (1985a).

⁹The converse is not true; it is possible to construct examples in which the algorithm fails to produce a matching even though there is a unique stable matching.

The corollary confronts us squarely with a puzzle. Typically there may be many stable outcomes to this kind of two-sided matching market, but the PBS algorithm is rarely observed to fail. To see how these two observations may be resolved, we will examine data from a number of rushes in Section V; but first consider the operation of the PBS algorithm as part of the larger market in which sororities may be able to admit more than q new members, even though they are not allowed to fill more than q positions through the algorithm. Specifically, consider a market in which the membership of each sorority S_k before formal rush is m_k , in which quota for the PBS algorithm is q , and in which the total allowable size for any sorority is T . Then, the number of positions sorority S_k may be able to fill either through formal rush or the informal rush which follows is $q_k = \max\{q, T - m_k\}$. That is, every sorority has the right to fill up to q positions (whether or not this will bring membership above T), and any sorority that has not filled q positions or does not have T members at the end of formal rush is able to continue recruiting new members. We have the following result.

THEOREM 3: *In the market with quotas q_k , matchings produced by the PBS algorithm with quota q may not be stable.*

PROOF:

Suppose the first $q + 1$ rushees on the bid list of some sorority S_k with $q_k > q$ all list S_k as their first choice. Further suppose that the $(q + 1)$ th rushee on S_k 's bid list, r_{q+1} , lists sorority S' as her second choice and that S' lists r_{q+1} among its first q rushees. Then, if the PBS algorithm with quota q results in a matching μ , Theorem 1 implies that $\mu(r_{q+1}) = S'$; but μ is unstable in the market with quotas q_k , since in that market S_k has a vacant position, and μ is blocked by (S_k, r_{q+1}) .

There is ample reason (both empirical and theoretical) to believe that instabilities give agents strong incentives to circumvent the procedures that produce them. Therefore, Theorem 3 raises a further question

about how the PBS algorithm has survived for so long. The empirical observations reported next will shed some light on this.

V. Some Empirical Observations

Preference cards and bid lists from formal rush were solicited from 12 campuses. These are regarded as highly confidential, and only four of the campuses agreed to make this material available, and then only under the condition that neither the names of sororities and rushees nor the campuses themselves would appear in any report. The data from 21 recent PBS-algorithm assignments taken from the four campuses are summarized in Tables 2, 3, and 4.¹⁰

Campus A is a rural college with approximately 1,500 full-time students; B is an urban university with a full-time undergraduate enrollment of about 4,500; C is a university in a rural setting with approximately 10,400 full-time undergraduates; and D is an urban university with roughly 9,400 full-time undergraduates. These four campuses are not a representative sample. All are located in the northeastern United States, and each had many sororities whose membership was sufficiently below their maximum capacity (their "total") so as to pose only loose constraints on the number of bids they could issue after formal rush.

As this latter factor will play a role in our subsequent analysis, the number of sororities that have "constrained" and "unconstrained" totals is noted below each table. Operationally, a sorority was said to be constrained only if the number of rushees on its second bid list who listed that sorority as their first choice was greater than the number of positions the sorority had available after formal rush (see Section VI). Other-

¹⁰In all but one case, the reported statistics are based upon the original preference lists (these were not available in 1986 on campus C: the statistics for that year were compiled by an administrator with direct access to the data). Some of the campuses retained old records and had many past PBS assignments available. Others only kept the most recent PBS assignments.

TABLE 2—CAMPUS A DATA

Statistic	1982 ^a	1983	Spring 1984	Fall 1984 ^a	1987
Number of rushees	93	84	72	92	68
Number of suicides	47	38	49	54	47
Matched	38 (41)	33	34	51 (46)	38
Unmatched	9 (6)	5	15	3 (8)	9
(Not listed)	0	5	4	0	0
Number with two choices	46	44	23	38	21
Matched	44 (43)	40	22	38 (37)	21
First choice	34 (35)	38	18	37 (35)	18
Second choice	10 (8)	2	4	1 (2)	3
Unmatched	2 (3)	4	1	0 (1)	0
(Not listed)	9	11	3	0	0
Number with three choices	0	2	0	0	0
Matched ^b	0	2	0	0	0
Unmatched	0	0	0	0	0
(Not listed)	0	4	0	0	0
Total matched (percentage)	88.2 (90.3)	89.3	77.8	95.7 (90.2)	86.8
First choice	77.4 (81.7)	86.9	72.2	95.6 (88.0)	82.4
Second choice	10.8 (8.6)	2.4	5.6	1.1 (2.2)	4.4
Total unmatched (percentage)	11.8 (9.7)	10.7	22.2	3.3 (9.8)	13.2
Suicides (percentage)	50.5	45.2	68.1	58.7	69.1

Notes: The maximum chapter size (T) was 55. Of the five sororities on campus A, two were constrained and three were unconstrained. There were two periods of formal rush in 1984. In 1984, the timing of formal rush was changed from spring (1982–1984) to fall. The 1985 fall formal rush results were unavailable. In 1986, formal rush was again changed to the spring.

^aIn 1982, an error occurred in the execution of the PBS algorithm. In fall of 1984, quota was incorrectly determined to be 25 when it should have been 21.8 or 22. The numbers shown in parentheses are the correct statistics based upon the correct assignments. All statistical tests are based upon the statistics resulting from the actual (not the correct) assignments.

^bBoth rushees who listed three choices matched to their first choice. Note that neither rushee was listed by her second- or third-choice sorority.

wise, the sorority was said to be unconstrained. With one exception, the constrained status of each sorority has remained unchanged over the years under observation.¹¹

The most striking feature of the data is the high percentage of rushees who chose to

list only one sorority on their preference card. This is particularly striking in view of the fact that this practice ("suiciding") is explicitly discouraged.¹² Nevertheless, of the

¹¹In the most recent PBS assignment occurring on campus D (1987), some sororities became constrained (see notes to Table 4).

¹²The following suggestions or guidelines were listed in an orientation booklet distributed to rushees during formal rush on one of the campuses we studied:

...if a rushee does not receive her first choice, she must be willing to accept any of the other choices she has listed. However, if she only preferences one sorority (sometimes called

TABLE 3—CAMPUS B DATA

Statistic	Spring 1979	Fall 1980	Spring 1981	1982	1983	1984	1985	1986	1987
Number of rushees	62	70	57	82	91	76	102	96	125
Number of suicides	53	58	53	79	67	51	86	73	80
Matched	52	51	38	56	57	47	75	50	63
Unmatched	1	7	15	23	10	4	11	23	17
(Not listed)	1	4	0	9	4	3	3	11	8
Number with two choices	9	11	4	2	23	24	15	20	34
Matched	9	11	3	1	21	23	13	18	31
First choice	9	10	3	1	15	19	10	13	23
Second choice	0	1	0	0	6	4	3	5	8
Unmatched	0	0	1	1	2	1	2	2	3
(Not listed)	1	2	1	0	9	6	4	8	17
Number with three choices	0	1	0	1	1	1	1	3	11
Matched	0	1	0	1	1	1	1	2	8
First choice	0	1	0	1	0	0	1	2	7
Second choice	0	0	0	0	0	1	0	0	1
Third choice	0	0	0	0	1	0	0	0	0
Unmatched	0	0	0	0	0	0	0	0	3
(Not listed)	0	0	0	0	1	1	0	4	14
Total matched (percentage)	98.4	90	71.9	70.7	86.8	93.4	87.3	72.9	81.6
First choice	98.4	88.6	71.9	70.7	79.1	86.8	84.3	67.7	74.4
Second choice	0	1.4	0	0	6.6	6.6	2.9	5.2	7.2
Third choice	0	0	0	0	1.1	0	0	0	0
Total unmatched (percentage)	1.6	10	28.1	29.3	13.2	6.6	12.7	27.1	18.4
Suicides (percentage)	85.5	82.9	93	96.3	73.6	67.1	84.3	76	64

Notes: The maximum chapter size (T) was 50. Of the six sororities on campus B, two were constrained and four were unconstrained. This campus had two formal rush periods every year, fall and spring, until 1982. The 1982 data represent the first year that there was only one formal rush period, held in the spring. Formal rush has continued to be held in the spring since 1982. There are two missing observations: spring 1980 and fall 1981. In 1986, an error occurred in the execution of the PBS algorithm. This error had no effect upon the aggregated statistics.

21 rushes observed on four campuses, there were only three in which the number of rushees suiciding was less than 50 percent of those who submitted preference cards. Even on campuses C and D, which each have a dozen or more sororities active in formal rush, relatively few rushees list more

than two sororities on their preference cards.

In Tables 2–4, the number of “suicides” is given immediately below the number of rushees submitting preference cards (with the percentage given in the last line of the table). For comparison, the number of rushees listing two choices on their preference lists and the number with three choices are also given. For each of these, the table shows how many were matched by the algorithm, broken down into how many are matched to their first, second, and third choices. Also shown for each category of rushees is the number of times a rushee

“suiciding”) she must realize she is limiting her chances of pledging a sorority all together.

No sorority shall encourage a rushee to single preference their sorority (suicide).

TABLE 4—CAMPUS C DATA AND CAMPUS D DATA

Statistic	Campus C			Campus D			
	1984	1985	1986	1984	1985	1986 ^a	1987
Number of rushees	59	79	93	89	78	96	119
Number of suicides	35	41	44	34	54	57	57
Matched	32	36	30	22	38	44	48
Unmatched	3	5	14	12	16	13	9
(Not listed)	0	0	1	0	0	0	0
Number with two choices	24	38	49	52	16	31	44
Matched	23	38	49	47	16	29	40
First choice	20	37	46	42	15	25 (24)	30
Second choice	3	1	3	5	1	4 (5)	10
Unmatched	1	0	0	5	0	2	4
(Not listed)	0	0	0	4	0	3	4
Number with three choices	0	0	0	3	8	8	18
Matched	0	0	0	3	8	8	17
First choice	0	0	0	3	7	6	13
Second choice	0	0	0	0	1	1	4
Third choice	0	0	0	0	0	1	0
Unmatched	0	0	0	0	0	0	1
(Not listed)	0	0	0	4	8	7	18
Total matched (percentage)	93.2	93.7	85.0	80.9	79.5	84.4	88.2
First choice	88.1	92.4	81.7	75.3	76.9	78.2	76.5
Second choice	5.1	1.3	3.2	5.6	2.6	5.2	11.7
Third choice	0	0	0	0	0	1.0	0
Total unmatched (percentage)	6.8	6.3	15.1	19.1	20.5	15.6	11.8
Suicides (percentage)	59.3	51.8	47.3	38.2	69.2	59.4	47.9

Notes: The maximum chapter size (*T*) was 65 on campus C and 55 on campus D. All of the 13 sororities on campus C and all of the 12 sororities on campus D were unconstrained during 1984–1986; in 1987, three of the sororities on campus D were constrained, and nine were unconstrained. Campus C requires that a sorority list all rushees who were extended a bid to its final party somewhere on its bid list. All rushes for both campus C and D take place in the fall. Quota-plus was adopted during the 1984 formal rush on Campus D, quota-only was adopted for all other years (1985–1987). On Campus D, sorority totals became relevant for the first time in 1987. The PBS algorithm failed to assign all rushees in 1987.

^aAn error occurred in the execution of the PBS algorithm on Campus D in 1986. The numbers shown in parentheses are the correct statistics based upon the correct assignments. All statistical tests are based upon the statistics resulting from the actual (not the correct) assignments.

placed on her preference card a sorority that did not in turn list the rushee, either on the first or second bid list.

For example, Table 3 shows that, in the 1979 rush on campus B, 62 rushees signed preference cards. Of these, there were 53 rushees who listed only a single sorority, and 52 of these were matched to that sorority, while one was unmatched. In this case, we can see that this is because one rushee

who listed only one sorority was not listed on that sorority's bid list. Similarly, nine rushees listed two sororities, and all nine were matched to their first choice. One of the sororities so listed did not place one of these rushees on either of its bid lists. Over 98 percent of the rushees were matched (all to the first choice on their preference cards), in a rush in which over 85 percent listed only one sorority on their preference card.

TABLE 5—DATA ON CONTINUOUS OPEN BIDDING

Campus	Year	Number of rushees unmatched during formal rush		Number of rushees matched to first choice during continuous open bidding	
		Suicides	Nonsuicides	Suicides	Nonsuicides
C	1985	5	0	1	0
C	1986	14	0	1	0
D	1984	12	5	11	4
D	1985	16	0	14	0
D	1986	13	2	13	2
D	1987	3 (9)	2 (5)	2 (4)	2 (3)

Notes: The 1984 data on campus C were unavailable. The 1987 results for campus D are broken into two groups, depending on whether the first-choice sorority was constrained or unconstrained. The first number is the number of rushees whose first choice was unconstrained, while the number in parentheses is the total number of rushees (i.e., those whose first choice was constrained plus those whose first choice was unconstrained). For example, nine rushees listed only a single sorority, of which three listed an unconstrained sorority. Of these three, two matched to their first choice in continuous open bidding.

On this campus, the maximum membership allowed in a sorority (T), which is the same for each sorority on campus, is 50, and of the six sororities on campus, only two were near enough this number so that it could constrain their post-PBS bidding.

The maximum number of rushees each sorority can be assigned under the PBS algorithm (quota) will vary each year. Quota is the number of rushees attending the first round of invitational parties divided by the number of sororities on the campus. Quota is not shown in the tables nor can it be calculated from the information given. The "number of rushees" shown in the tables is the number signing preference cards, which may be substantially smaller than the number of rushees attending the first round of preference parties.

The PBS algorithm failed to assign all rushees (as either matched to a sorority or as unmatched) on campus D during the 1987 formal rush. This was the only failure observed. Those rushees not assigned by the PBS algorithm were assigned by the individual in charge of the execution of the PBS algorithm. (The resulting matching was stable; see Mongell [1988] for an analysis of this incident.) The statistics in Tables 2–4 indicate the assignments made by the PBS algorithm. The quota-only method was

adopted by all the campuses observed, except for one year (1984) on campus D. Under quota-only, sororities may extend additional bids to rushees assigned as unmatched by the PBS algorithm. The procedure by which these additional bids are extended varies on each campus (and sometimes from year to year). The sorority may be notified that an unmatched rushee's name appears on its bid list and asked whether it would like to extend her a bid. The sororities may be notified before the PBS execution that unmatched rushees whose names appear on their bid list will be extended bids (on the sorority's behalf) if the sorority has not reached quota during the PBS algorithm. From the available data it was observed that few sororities extended bids at this time. On some campuses, a rushee assigned as unmatched by the PBS algorithm will be called by one of the individuals involved with the PBS execution and asked whether she would be willing to join another sorority that listed her on its bid list and has not reached quota. It appears from the (limited) available evidence on this point that virtually all rushees so called have refused these bids.

The numbers of rushees assigned as unmatched by the PBS algorithm who match to their first choice during continuous open

bidding were available on campuses C and D and are shown in Table 5.

VI. Strategic Analysis

The data raise two questions. What accounts for the consistently high percentage of rushees who list only a single sorority on their preference cards? And how might this high percentage be related to the low frequency of failure of the PBS algorithm and to its long life? To address these questions requires a model of the entire rush process.

During the PBS algorithm, each sorority may gain up to q new members. Following the PBS algorithm, there is a second stage of formal rush during which one additional set of bids and acceptances or rejections may be made (with how many bids depending on whether the quota-only or quota-plus rules are adopted). Finally, following formal rush there is continuous open bidding, during which each sorority with fewer than T members (both new members and old members who have not yet graduated) may admit new members to bring its membership up to T (and each sorority that has not yet enrolled q new members may bring its new members up to q). On the campuses from which our data are drawn, T imposed such a loose constraint that most sororities could attempt to recruit all rushees who showed serious interest in them. (These are the "unconstrained" sororities noted below in Tables 2–4.) That is, on these campuses, the demand for membership in most sororities is less than the supply, in the sense formalized by the following definition.

Definition: A sorority S is unconstrained if its membership is sufficiently below its allowed total so that it can extend bids to all rushees who it finds acceptable and who have S as their first choice among all sororities that find them acceptable.

Nevertheless, the number of rushees interested in joining even an unconstrained sorority may exceed q . As we saw in the proof of Theorem 3, a rushee who lists more than one sorority on her preference card runs the risk of being matched to her

second-choice sorority during the PBS algorithm and forgoing a chance to be matched to her first-choice sorority after the formal rush. For simplicity, consider the case in which all sororities have unconstrained totals (i.e., for every sorority S_k , the number of acceptable rushees who regard S_k as their first choice is less than q_k , where $q_k = \max\{q, T - m_k\}$ is the number of positions S_k may fill by the end of open bidding).

We model the matching procedure as a multistage game. In stage 1, all sororities and rushees simultaneously state preferences and are matched by the PBS algorithm. In stage 2, the unmatched rushees are announced, as are the sororities that have not filled q positions (if the quota-only rules are used, or which have fewer than T members, if the quota-plus rules are used). Each such sorority S_k may issue invitations to up to $q - |x(S_k)|$ (or $q_k - |x(S_k)|$) unmatched rushees, and each rushee who receives invitations may accept at most one. In stage 3 and subsequent stages, all matches from previous stages become public, and any sorority S_k that has been matched to a set $y(S_k)$ of rushees in the prior stages may issue invitations to up to $q_k - |y(S_k)|$ rushees who have not been matched to sororities in earlier stages. Each rushee may accept at most one invitation and must decline all others when they are received; at any stage in which she accepts an invitation, she is matched. Starting with stage 4, no sorority may issue an invitation to a rushee to whom it has previously issued an invitation at stage 3 or later. The game ends at any stage in which no invitations are issued. Stages 1 and 2 represent formal rush, with stage 1 corresponding to the PBS algorithm and stage 2 corresponding to the quota-only (or quota-plus) system. Subsequent stages represent open bidding.

Note that we have chosen one of several possible ways to model the second stage of formal rush. Also, by dividing the bidding into stages we have imposed on the model some structure beyond what we observe in practice in open bidding. Finally, so the game will end in finitely many periods, we have imposed the rules that sororities may

not reinvoke rushees and that rushees must either accept or reject all invitations in the period they are received. These choices seem to be among the simplest that are broadly consistent with the sometimes diverse and sometimes ambiguous rules of the rush process. Because there is some irreducible arbitrariness in choosing the elements of a model, it is also important to note that the equilibrium considered below seems robust to changes in these arbitrary features of the model.

We would like to demonstrate that the observed behavior corresponds to equilibrium behavior in this market. One potential difficulty we face is that we have not fully specified what happens when the PBS algorithm fails. To show that a particular set of strategies is in equilibrium, we have to show that no agent can profitably deviate, and for this we have to show that no agent can profitably deviate even in a way that causes the algorithm to fail. As noted earlier, different individuals charged with supervising sorority rush have indicated that they would proceed differently in the circumstances we call failure (i.e., in these circumstances, the results would be different on different campuses). One approach, therefore, would be to make further assumptions about how the algorithm would proceed on each of these campuses. We take a different approach and demonstrate an equilibrium with the properties that only rushees can deviate in a way that might cause the algorithm to fail and no rushee can profit from this, no matter how failures are resolved (however failures of the PBS algorithm might be resolved on different campuses, rushees may not be matched to sororities that have not issued them invitations).

THEOREM 4: *Suppose all sororities are unconstrained. Then:*

(a) *The following strategies constitute an equilibrium in the multistage game. In the first stage, each rushee lists only her first-choice acceptable sorority (from among those that find her acceptable) and in subsequent stages accepts only an offer from this sorority. Each sorority S_k lists its true preferences in the first stage and makes offers in stage 2 to*

its most preferred $q_k - |x(S_k)|$ acceptable rushees from among those unmatched in stage 1. In stage 3, it makes offers to all of its $(q_k - |y(S_k)|)$ highest-ranked unmatched rushees, and in any subsequent stage it makes offers to its highest-ranked set of unmatched rushees who have not previously rejected it.

(b) *Furthermore, at this equilibrium the PBS algorithm never fails, and the matching that results is stable in the market with quotas q_k (it is the rushee-optimal stable matching).*

PROOF:

We prove part b first. Suppose the rushees and sororities play the strategies described. To see that the PBS algorithm does not fail, suppose to the contrary that it ends with some rushee r_i in hold. Then r_i must be on the second bid list of the sorority on the top of her preferences at the final step of the algorithm, and this sorority, S , must not have reached quota (since if it had it would have been crossed off r_i 's preference list at box D of the flow chart in Fig. 1). Therefore, there must be another rushee, r_j , not matched to S but in the first q positions of S 's final bid list. However, this cannot be, since r_j has listed only one sorority. If this is not S , then r_j would have been crossed off S 's list at box B of the flow chart, and if it is S , then (all such) r_j would be matched to S , contradicting that S has not reached quota. Therefore, the PBS algorithm leaves no rushees in hold; that is, it does not fail.

When all agents play these strategies, each rushee is eventually matched to her first choice among all acceptable sororities that find her acceptable. Since this matching is individually rational and not blocked by any sorority-rushee pair, it is stable, and since each rushee is matched to her highest-ranked achievable match, it is the rushee-optimal stable matching μ_R . This completes the proof of part b.

To prove part a, we show that no sorority or rushee can do better than to play the strategy described, so long as the other agents all do so. First consider sororities. As we saw in the proof of part b, so long as all rushees list only a single sorority on their preference cards, the algorithm will not fail, regardless of what sororities may do. Fur-

thermore, if the rushees follow the strategies indicated in the theorem, any sorority S that deviates from the strategy indicated for it will be matched to a subset of $\mu_R(S)$, rather than to all of $\mu_R(S)$. Since S has responsive preferences over groups of rushees, and since all rushees in $\mu_R(S)$ are strictly preferred to vacant positions, S prefers $\mu_R(S)$ to any strict subset of $\mu_R(S)$, and so cannot profit from any such deviation.

Now consider rushees. For each r , $\mu_R(r)$ is r 's most-preferred mutually acceptable sorority; that is, $\mu_R(r)$ is the most preferred match r can achieve at any individually rational outcome (if there are no mutually acceptable sororities, $\mu_R(r) = r$.) Therefore, if r deviates from her indicated strategy, she cannot improve her outcome even if by deviating she causes the PBS algorithm to fail, since no rushee may be matched to a sorority that has not issued her an invitation.

While the equilibrium specified in Theorem 4 is not a perfect equilibrium, the equilibrium behavior is certainly consistent with perfectness. Since the extensive-form game begins with the simultaneous submission of all parties' preferences, all equilibria are subgame perfect. However, after formal rush, all parties learn all the payoff-relevant information of the game, and the subsequent information sets all consist of single nodes, so an appropriate formulation of perfectness is backward induction to the nodes of stage 3. The off-the-equilibrium-path behavior we must consider arises if a rushee's first-choice sorority fills all its positions before issuing her an invitation. In this event, the rushee's strategy should be, from stage 3 onward, to accept the offer from her highest-ranked sorority among those that will still have positions to offer when they reach her, following the strategy for sororities given in the theorem.¹³ Note also that

the stage-2 behavior of sororities plays little role in this equilibrium (e.g., nothing would change if sororities made no offers in stage 2 but otherwise behaved as in Theorem 4).

Theorem 4 considers the case in which all sororities are unconstrained, whereas in our data this was the case only on campuses C and D; both campuses A and B had some constrained sororities, although most were unconstrained. Thus, the assumptions of the theorem do not precisely model the situation we observed any more than the equilibrium strategies it characterizes precisely mirror the data, which on every campus show significant numbers of rushees listing more than a single sorority on their preference cards in almost every year. Similarly, the equilibrium outcome has all rushees ultimately matched to the sorority they list first on their preference card, while Table 5 shows that, of campuses C and D, this approximately characterizes only the situation on campus D. However the theorem shows how the striking regularities observed in our data can arise at equilibrium. It shows how stage 2 of the formal rush procedure plays a much less important role than does the continuous open bidding which follows formal rush. Most importantly, it makes clear why the presence of unconstrained sororities may be expected to give so many rushees an incentive to list only a single sorority. Even when some sororities have tighter constraints, this incentive persists, since for example a rushee whose first-choice sorority is constrained but whose second-choice sorority is unconstrained also has no incentive to list more than her first choice when the strategies are as described.

Theorem 4 and its proof also suggest why an increase in the number of rushees who list only one choice on their preference cards will reduce the probability that the PBS

¹³If the constraints on sororities were completely relaxed (e.g., if the quotas q_k all exceeded the number of rushees), then the equilibrium in Theorem 4 would be perfect, and μ_R would be the unique stable match-

ing. However such a relaxed constraint does not describe what we observed. Similarly, we could have modeled open bidding by rules that would lead to stable outcomes at every perfect equilibrium regardless of the quotas (see Roth, 1984b; Roth and Sotomayor, 1990), but this would involve imposing particular detailed rules beyond what we observe.

algorithm will fail, even if some rushees behave differently. That is, increasing the number of rushees who submit a single choice on their preference cards may remove the cause of failure of the PBS algorithm but may never cause failure. The following proposition, stated without proof, formalizes this.

PROPOSITION: *Let P be a collection of stated preferences for a set S of sororities and a set R of rushees, and let P' be a collection that differs from P only in that some of the preference orderings in P have been truncated after their first element. Then the PBS algorithm with input P' will never fail if the PBS algorithm with input P does not.*

VII. Modeling Issues and Open Questions

Many choices must be made in modeling a complex system. We have already pointed out some of the modeling decisions we have made. Here, we discuss some aspects of sorority rush that we have not included in the formal analysis. Our motivation for discussing these explicitly is that, if such choices are not made carefully, the conclusions of the analysis may be misleading. Therefore, we want to explain briefly the reasons behind our choices. Then, we turn to some open questions.

First, although we have modeled sororities as being concerned with groups of rushees, we have modeled rushees as having preferences only over sororities and not over which other rushees join the same sorority. This seems justified both because particular sororities typically draw from the same part of the rushee population year after year (so preferences over sororities are a good proxy for preferences over other rushees) and because rushees are typically freshmen who have not yet had time to form many close friendships with other freshmen. Nevertheless, it is not unheard of for pairs of rushees, typically friends from high school, to wish to join the same sorority, and the problem facing such a pair differs from that analyzed here.

Second, our strategic analysis considered only the behavior of individual rushees and

sororities and not sorority-rushee coalitions. There may be an additional reason why some rushees list only a single sorority, since in some circumstances it may be in the interest of an unconstrained sorority to encourage certain rushees to do so, although this is regarded as one of the more serious violations of the rules. Briefly, certain rushees (called "legacies") may have close relations with a given sorority even before the beginning of rush, by virtue of having a family member who is a member or alumna of that sorority. If this rushee r lists only that sorority S on her preference card, then sorority S can plan to list rushee r somewhere on its second bid list, and can count on enrolling her during open bidding. This permits the sorority to rank higher on its bid list other rushees, who may list more than one sorority and who might therefore be matched during the PBS algorithm to another sorority if sorority S submitted its true preferences. We have been unable to gather data on how widespread this phenomenon might be, both because it is explicitly forbidden by the authorities concerned with rush and because it is not easily distinguishable from the simpler reasons for listing only a single sorority already described.¹⁴ That is, the reasons this might be a viable agreement between a sorority and a rushee are not substantially different from the reasons that individual rushees, acting on their own, might choose to submit a single preference.

Third, we have not analyzed the several rounds of parties described in Section II, which precede the submission of preferences by sororities and rushees. Our sense is that, on the campuses we have observed, because most sororities are unconstrained or only loosely constrained, the strategic considerations that arise in deciding which rushees to invite and which parties to accept have at most secondary importance, and the primary role of the parties on these

¹⁴We are indebted to Patty Beeson for pointing out to us that some sororities have rules that any legacy who attends the final preference party must be listed on the first bid list.

campuses is to help rushees and sororities form their preferences and signal them to one another.

This brings us naturally to the next modeling issue. We have analyzed the game as a game of complete information, in which sororities and rushees know one another's preferences. Essentially we are assuming that, in the course of the preference parties, these preferences are fully communicated. This is obviously only an approximation of reality, but a rough idea of how adequate this approximation might be can be gotten from Tables 2-4, which show how many times a sorority was listed on the preference card of a rushee who did not appear anywhere on its bid list. Each such incident is likely to be a case in which the complete-information assumption is not met.¹⁵ The relatively low frequency of this suggests that the complete-information assumption is a rough approximation of what we observed. Note that, while the assumption of complete information is certainly less than fully satisfactory, serious new problems would arise in attempting to model the game as one of incomplete information, since the results of such an analysis would be sensitive to the assumptions that would have to be made about participants' prior probability distributions.¹⁶

Finally, our analysis has treated each sorority as an individual agent and not as a collection of individual members. Given that each sorority is required to submit a single

bid list, this may not raise problems, but we note that we have not investigated any aspect of how sororities arrive at their bid lists. In this regard, Roth and Sotomayor (1989) show there is a surprising coincidence of preferences over stable matchings among agents with different responsive preferences over groups, provided they have the same preference over individuals. Thus, for many purposes, the relevant differences among sorority members will be precisely those that go into determining the preferences over individuals (i.e., the bid list) and not more complex issues regarding the makeup of the entire entering group of new members.

We now consider briefly one of the major open empirical questions raised by this work: on campuses having mostly constrained sororities, how will rush differ from what we have observed on campuses with mostly unconstrained sororities? We conjecture that there will be at least two important (and related) differences. As we have seen, the high percentage of rushees listing only one sorority on their preference cards in formal rush is related to the fact that this (unconstrained) sorority can issue further invitations during open bidding.¹⁷ This will not be so on campuses in which most sororities cannot accept new members after the end of formal rush, and so on these campuses we expect to see a very much smaller percentage of single preferences.

This brings us to the second major difference we expect. Recall that the PBS algorithm as delineated in the literature of the National Panhellenic Council is incompletely specified: for some configurations of preferences, it does not indicate how some rushees should be dealt with. As we saw, the very low frequency of this kind of failure in our data can be attributed to the high percentage of rushees who submit single preferences. If the percentage of single preferences is much lower on campuses with

¹⁵Of the four campuses observed, only the sororities on campus C are required by their college panhellenic to list every rushee invited to the final preference party somewhere on their bid list. Thus, the invitations to the final round of parties are particularly effective signals of sorority preferences on campus C (but even on this campus, there was one case of an unlisted rushee in 1986; see Table 4).

¹⁶Although perhaps some headway could be made due to the fact that the preferences of rushees for sororities and those of sororities for rushees may follow certain identifiable patterns (e.g., on a given campus some sororities may be known as athletic, others as wealthy, etc.). A discussion of equilibria when agents have incomplete information about other agents' preferences is found in Roth (1989) and Roth and Sotomayor (1990).

¹⁷At least in theory. The differences between the various campuses in our data set (see the notes in Tables 2-4) preclude meaningful statistical comparisons across campuses (see Mongell, 1988).

constrained sororities, it seems likely that the frequency of failure (i.e., the percentage of rushes in which the submitted preferences fall outside of those that can be fully processed by the standard instructions for the PBS algorithm) will be higher. (Indeed, the one failure that occurred in our data was on campus D in 1987, the first year some of the sororities on that campus became constrained.) This suggests that supplementary rules will be adopted on these campuses to determine what the algorithm should do in such cases. Since it appears that these rules will have to be developed separately on each campus, there may be more variation in the formal rush procedures found on such campuses (as well as in the strategic behavior of rushees and sororities).

Finally, on campuses with many constrained sororities, it seems likely that the initial rounds of preference parties would involve nontrivial strategic decisions. (For example, if you can only accept two final invitations, it might sometimes be advisable to decline an invitation from your first-choice sorority, in order to signal your interest to a lower-ranked choice which has a greater chance of giving you a high ranking on its preference list.) As described in Section II, a good deal of formal structure is involved in these parties, much of which is common from campus to campus, which would help in modeling the strategic aspects of these decisions.¹⁸

A second kind of open question is: what caused the unraveling of recruitment dates that occurred in this market prior to the introduction of the PBS, and what caused this unraveling to stop? Sorority rush may not be the two-sided matching market that will best illuminate these issues,¹⁹ but be-

cause this phenomenon occurs in other two-sided matching markets, the unraveling observed in sorority rush appears to be an example of a much more general phenomenon (see Roth 1984a, 1991).

VIII. Conclusions

One way to summarize this story is to say it is about the difficulty of central planning. The agreement hammered out in the 1920's among sororities to reduce the competitiveness of recruitment implemented a plan designed to give all sororities the ability to recruit the same number q of new members, before any sorority had a chance to recruit more. When sororities have capacity greater than q , which is the case on the campuses we have observed, this could in general lead to unstable outcomes (Theorem 3). That is, there will in general be rushees and sororities who share an incentive to circumvent this constraint. As we have seen, the agents in the market have adapted their behavior to do so: rushees list only their first-choice sororities in the part of the procedure constrained by the quota q , and sororities approach desirable rushees after this constraint has been lifted. Ironically, this adaptation contributes to the smooth operation of what would otherwise be an incompletely specified procedure (Theorem 4).

From a more general perspective, this study reaffirms the importance of examining systems of rules from the point of view of what incentives they give participants. If we had looked only at the formal rules of the PBS algorithm and analyzed it as if agents all submitted their true preferences, we would not have been able to explain what we were seeing. However, the data on submitted preferences clearly made it unlikely that rushees were submitting their true full

¹⁸For example, of the campuses we observed, all but campus B permit each rushee to attend only two final parties during the last round of the invitational parties. (Campus B permits rushees to attend up to three final preference parties.)

¹⁹Since sororities are subject to some sanctions (both from the national organization and from campus authorities), they may be able to enforce an agreement on

recruiting behavior simply, once it has been reached. Also, with the increased mobility of college students, there may not be much room to unravel recruiting much before the beginning of the freshman year (i.e., the unraveling may have stopped only because it has no further to go).

preferences, and the subsequent analysis of their incentives suggested that they were responding rationally to the incentives of the system.

We reemphasize that we study sororities not merely because of their intrinsic interest, nor merely to show that game-theoretic analysis can shed light on behavior that might not always be thought of as economic. The point of studying particular matching mechanisms is that they add to our understanding of how centralized matching mechanisms work in general (and there seems to be a surprising number of these). By understanding how centralized mechanisms work in practice, we can also hope to learn things that will be useful in the study of decentralized two-sided matching markets (see Roth and John H. Vande Vate, 1990, 1991). The advantage of beginning with centralized markets is that it is easier to determine when they reach stable outcomes and when they do not. The theoretical progress in studying labor and other markets as two-sided matching models (see references in Roth and Sotomayor [1990]) suggests that this kind of empirical research may be fruitful. The conclusions of the present study should lend further weight to the hypothesis that the stability or instability of the matchings that result from such a market are crucial to understanding the market's evolution.

If game theory is to become as important a part of empirical economics as it has become a part of economic theory, we must explore the kinds of empirical research that will allow us to test and refine game theory. Since game theory is the part of economic theory most particularly concerned with the "rules of the game," this will inevitably involve looking into the particular rules by which markets are organized. At the same time, as was emphasized in Roth (1991), notions like the stability of outcomes, which can be formulated somewhat independently of the particular rules of the game, are useful in comparing different sets of rules and in indicating when agents may have an incentive to change the rules or circumvent them.

APPENDIX

The following results from the literature will be of use.²⁰

THEOREM A1: *A stable matching exists for every matching market.*

Definition: For a given matching market (S, R, P) , a stable matching μ is S-optimal if every sorority likes it at least as well as any other stable matching. Similarly, a stable matching ν is R-optimal if every rushee likes it at least as well as any other stable matching.

Definition: A sorority S and a rushee r are achievable for each other in a matching market (S, R, P) if S and r are paired at some stable matching.

In a matching market in which all rushees and sororities have strict preferences over individuals and in which sororities have responsive preferences over groups, each sorority and rushee can strictly order its or her achievable mates, and so there can be at most one S-optimal stable matching and one R-optimal stable matching.

THEOREM A2: *When all sororities have strict preferences over individual rushees and all rushees have strict preferences over sororities, there always exists an S-optimal stable matching, μ_S , and an R-optimal stable matching, μ_R .*

THEOREM A3: *When all agents have strict preferences, the S-optimal stable matching is the worst stable matching for all the rushees; similarly, the R-optimal stable matching is the worst for all the sororities.*

²⁰Theorems A1 and A2 were proved in Gale and Shapley (1962), and Theorem A3 was proved in Donald Knuth (1976) for the marriage model. They hold not only for the marriage model, but also for the college admissions model considered here (see Roth and Sotomayor, 1990). Theorem A4 is from Roth (1984a); for a stronger result motivated by the American medical labor market, see Roth (1986).

We will also use the following result.

THEOREM A4: *When all preferences over individuals are strict, the set of rushees matched to sororities, and sorority positions filled, is the same at every stable matching.*

Theorem 1 follows immediately from the following proposition, which will also be useful in the proof of the next theorem.

PROPOSITION 1: *If r_i is not in hold when the algorithm stops, then any sorority S_k that r_i prefers to $x(r_i)$ does not prefer r_i to any element of $x(S_k)$.*

PROOF:

Consider a rushee r_i who is not in hold when the algorithm stops. Let t be the step at which this rushee is given an assignment (i.e., either matched to a sorority or left unmatched). Then $x'(r_i) = S_i$ for some sorority S_i , or $x'(r_i) = r_i$. In the former case (box A in Fig. 1), S_i is at the head of rushee r_i 's preference card at step t . (Note that such an assignment must be individually rational for both r_i and S_i .) In the latter case (box E or F), either the rushee's current preference card is empty (box F) or it contains only a sorority S_j that did not list rushee r_i (box E).

There are two points in the algorithm at which a sorority S_k can be deleted from the rushee r_i 's preference card: these are boxes C and D in Figure 1. If the deletion occurs at box C, the sorority has not listed rushee r_i on its bid list and so does not prefer rushee r_i to any element of $x(S_k)$. If the deletion occurs at box D, then the sorority has filled its quota by matching to q rushees at the top of its bid list during some step $k < t$. Since rushees are deleted from the bid lists of sororities only at boxes A and B, rushee r_i has not been deleted from sorority S_k 's bid list prior to step k , so S_k prefers all the rushees in $x(S_k)$ to r_i .

The next proposition states that, even when the PBS algorithm fails to assign all rushees, the resulting partial matching could be extended to a stable matching in the market with quota q .

PROPOSITION 2: *There exists a stable matching μ in the market with quota q such that $\mu(r) = z(r)$ for every rushee r who is matched by the PBS algorithm.*

PROOF:

Consider the "residual matching market" which arises in the market with quota q after the PBS algorithm has ended in failure, with some rushees left in hold. Define the agents in this market to consist of those rushees left in hold, together with the set of sororities. The preferences of the sororities are as in their bid lists in the original market, except that all rushees who have been matched by the PBS algorithm are deleted. The residual quota of a sorority S_k in the residual market is given by $q'_k \equiv q - |x(S_k)|$. That is, in the residual matching problem we have just defined, each sorority may fill no more positions than were left unfilled by the PBS algorithm. (The preferences of the rushees in this residual matching problem can be thought of either as the same as given by their preference cards in the original problem or as having deleted those sororities that now have a quota of 0.)

Let x' be a stable matching for the residual matching problem. Let μ be the matching for the original matching problem with quota q , such that for each rushee matched by the PBS algorithm, $\mu(r) = x(r)$ and for each rushee left in hold, $\mu(r) = x'(r)$. We will show that μ is stable in the market with quota q .

Suppose not. Then, since μ is individually rational, there is a sorority rushee pair (S, r) such that r prefers S to $\mu(r)$, and S prefers r to some σ in $\mu(S)$. It cannot be that r was assigned by the PBS algorithm, since if that were the case we would have $\mu(r) = x(r)$, which implies that, at the step of the algorithm at which r was matched to $x(r)$, sorority S had already filled its quota, so $\mu(S) = x(S)$, and by Proposition 1, (S, r) is not a blocking pair. Therefore, r was left in hold at the end of the PBS algorithm. Then, σ cannot be equal either to S (an unmatched position) or to a rushee r' in the residual matching problem, since this would contradict the stability of x' in that market. Therefore, σ must equal some rushee r'

with $x(r') = S$. Note that r cannot have listed only one sorority on her preference card, and so the fact that r' was matched to S while r was still unmatched implies $|x(S)| < q$ (since rushees can be deleted from sorority bid lists only at points A and B in the flow chart in Fig. 1). Therefore, r is one of the q' highest rushees in S 's preferences in the residual market. Thus, x' cannot be stable in the residual market, which is the contradiction needed to conclude that no blocking pair (S, r) exists, and so μ must be stable.

LEMMA: *If S_k is r_i 's most preferred achievable match and r_i is among S_k 's q most preferred achievable matches, then (r_i, S_k) is an inevitable match in the market with quota q .*

PROOF:

Theorem A3 implies that S_k is both the best and the worst achievable sorority for r_i . Since preferences are strict, it is the only achievable match.

PROOF OF THEOREM 2:

The proof is by induction. Bear in mind that, by Proposition 2, if r_j is matched to S_p at some step of the algorithm, r_j and S_p are achievable for one another.

If, when $t = 1$, $r_i \in Q_1(S_i)$ and if, of the sororities that listed r_i , S_j is the highest-ranked sorority on r_i 's preference card, then $x^1(r_i) = S_j$, and this first match is inevitable. (In what follows, the ranking on the rushee's preference card refers to the original ranking when $t = 0$.) Thus, if a match is made at $t = 1$, it is inevitable. Now suppose that all matches up to time $t = k$ are inevitable; it will be shown that any match made at $t = k + 1$ is inevitable.

Consider the following cases:

(i) No match occurs at step $k + 1$.

Then all matches up to step $k + 1$ are inevitable by the inductive assumption.

(ii) A match $x^{k+1}(r_j) = S_p$ occurs at step $k + 1$ (i.e., rushee r_j is matched to sorority S_p at that step).

Then it must be that $r_j \in Q_{k+1}(S_p)$ and that, of the sororities that listed r_j and have not reached quota, S_p is the highest-ranked sorority on r_j 's preference card. First suppose that S_p is the highest-ranked sorority on r_j 's (original) preference card. Then the fact that $x^{k+1}(r_j) = S_p$ implies that S_p has not reached quota prior to step $k + 1$. (a) If $r_j \in Q_0(S_p)$, then (r_j, S_p) is inevitable (by the Lemma). (b) If $r_j \notin Q_0(S_p)$ but $r_j \in Q_0^+(S_p)$, then there exist other rushees that S_p prefers to r_j . Since $r_j \in Q_{k+1}(S_p)$, this implies that all of these rushees have either suicided or matched elsewhere (since rushees are deleted from sorority bid lists only at boxes A and B of the PBS-algorithm flow chart in Fig. 1). Any rushee who suicided cannot be matched to S_p at any individually rational outcome. Any rushee who was matched to some sorority other than S_p prior to step $k + 1$ is matched to that sorority at any stable outcome, by the inductive assumption. Therefore, none of these rushees can be matched to S_p at any stable outcome, and (r_j, S_p) is inevitable, by the Lemma.

Now suppose S_p is not the highest-ranked sorority on r_j 's (original) preference card of those that listed her somewhere on their bid lists. Let S_n be a sorority that listed r_j and was ranked before S_p on her preference card. Then the fact that $x^{k+1}(r_j) = S_p$ implies that S_n has reached quota at some step $t < k + 1$ (since sororities are deleted from rushee bid lists only at boxes C and D of the flow chart in Fig. 1). By the inductive hypothesis, S_n is not achievable for rushee r_j . Therefore, by the Lemma, r_j and S_p are inevitable.

(iii) A match $x^{k+1}(r_j) = r_j$ occurs at step $k + 1$ (i.e., rushee r_j is assigned to be unmatched at that step).

Proposition 2 and Theorem A4 together imply in this case that $\mu(r) = r$ at every stable matching.

REFERENCES

Brown, James T., ed., *Baird's Manual of American College Fraternities*, 9th Ed.,

- New York: James T. Brown, 1920.
- Gale, David and Shapley, Lloyd, "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, 1962, 69, 9-15.
- Knuth, Donald E., *Marriages Stables*, Montreal: Les Presses de l'Université de Montreal, 1976.
- Mongell, Susan, *Sorority Rush as a Two-Sided Matching Mechanism: A Game-Theoretic Analysis*, Ph.D. Diss., Department of Economics, University of Pittsburgh, 1988.
- Roth, Alvin E., (1984a) "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory," *Journal of Political Economy*, December 1984, 92, 991-1016.
- _____, (1984b) "Misrepresentation and Stability in the Marriage Problem," *Journal of Economic Theory*, December 1984, 34, 383-7.
- _____, (1985a) "The College Admissions Problem Is Not Equivalent to the Marriage Problem," *Journal of Economic Theory*, August 1985, 36, 277-88.
- _____, (1985b) "Common and Conflicting Interests in Two-Sided Matching Markets," *European Economic Review* (special issue: Market Competition, Conflict, and Collusion), February 1985, 27, 75-96.
- _____, "On the Allocation of Residents to Rural Hospitals: A General Property of Two-Sided Matching Markets," *Econometrica*, March 1986, 54, 425-7.
- _____, "Two Sided Matching with Incomplete Information about Others' Preferences," *Games and Economic Behavior*, June 1989, 1, 191-209.
- _____, "A Natural Experiment in the Organization of Entry-Level Labor Markets: Regional Markets for New Physicians and Surgeons in the United Kingdom," *American Economic Review*, June 1991, 81, 415-40.
- _____, and Sotomayor, Marilda, "The College Admissions Problem Revisited," *Econometrica*, May 1989, 57, 559-70.
- _____, and _____, *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Econometric Society Monographs, Cambridge: Cambridge University Press, 1990.
- _____, and Vande Vate, John H., "Random Paths to Stability in Two-Sided Matching," *Econometrica*, 1990, 58, 1475-80.
- _____, and _____, "Incentives in Two-Sided Matching with Random Stable Mechanisms," *Economic Theory*, 1991, 1, 31-44.
- Shepardson, Francis W., ed., *Baird's Manual of American College Fraternities*, 12th Ed., Menasha, WI: Collegiate Press, 1930.
- National Panhellenic Conference, "How To" for College Panhellenics, 10th Ed., unpublished internal manual, 1979.
- _____, NPC: An Historical Record of Achievement, Maury Boyd, 1983 (unpublished facsimile of 1957 version published by Leland Publishers, St. Paul, MN).
- National Panhellenic Review, unpublished minutes, 1985.

Model Uncertainty, Learning, and the Gains from Coordination

By ATISH R. GHOSH AND PAUL R. MASSON*

This paper considers gains from international economic policy coordination when there is uncertainty concerning the functioning of the world economy but also learning about the "true" model on the part of policymakers. The paper reports estimates of plausible alternative versions of a standard two-country model. Activist policy (either coordinated or uncoordinated) may produce large welfare losses in the absence of learning, if policymakers believe in the wrong model; hence, exogenous money targets and freely flexible exchange rates may be best. However, model learning (from observations on macroeconomic variables) causes coordinated policies to dominate activist uncoordinated policies or exogenous money targets. (JEL F33, F42)

In recent years, there has been a marked increase in interest in international economic policy coordination, as evidenced by the proliferation of academic publications and of meetings of officials. Though the presumption is that policy coordination among the major industrial countries is a good thing, there exist valid doubts concerning the possibility of designing appropriate intervention and coordination rules when the effects of policies are uncertain. Speaking at the American Economic Association meetings in December 1987, Martin Feldstein noted, "Uncertainties about the actual state of the international economy and uncertainties about the effects of one country's policies on the economies of other countries make it impossible to be confident that coordinated policy shifts would actually be beneficial" (Feldstein, 1988 p. 10). Though coordinated policies may, *ex post*, turn out to have been ill-advised, the relevant question is whether they are likely to result in

higher welfare on average than uncoordinated policies, despite the presence of such uncertainties.

The issue has been discussed in several recent papers, though a consensus on the implications of model uncertainty for the desirability of coordination has yet to be achieved. Jeffrey Frankel and Katharine Rockett (1988) argue that model uncertainty makes coordination too risky and that, on average, countries are as well off pursuing noncooperative policies as they are under coordination.¹ In contrast, we have argued in another paper that the existence of model uncertainty does not necessarily preclude a beneficial role for the coordination of macroeconomic policies; indeed, it may in fact provide an additional incentive to coordinate policies internationally (Ghosh and Masson, 1988; see also Ghosh and Swati Ghosh [1991]). The reason for this is that, in the choice of optimal policies, coordination will properly take into account the effects of those policies on global economic uncertainty.

The differences in conclusions stem from two essential differences in approach. First

*Department of Economics, Princeton University, Princeton, NJ 08544 and International Monetary Fund, Washington, DC 20431. We are grateful to Mike Dooley, Dale Henderson, and an anonymous referee for helpful comments. The views expressed in the paper are those of the authors alone, and in particular do not represent those of the International Monetary Fund.

¹Gerald Holtham and Andrew Hughes Hallett (1987) show that criteria can be applied that diminish the likelihood that coordination will be bad.

is the question of whether to evaluate gains *ex ante* or *ex post*. Our earlier paper focuses on the *ex ante* expected gains from coordination, while Frankel and Rockett (1988) consider the *ex post* actual gains after arbitrarily specifying which is the true model. The second issue concerns the nature of expectations formation. In our earlier paper, policymakers are assumed to have rational expectations across the set of possible models and to take due account of the presence of model uncertainty in formulating policies, whereas in Frankel and Rockett's approach policymakers are assumed to have different subjective priors with respect to the different models, but each believes (wrongly) that he/she knows the correct model.²

In this paper, we reexamine the issue of model uncertainty and policy coordination, highlighting the effects of the differences in approach described above. We use variants of a simple consensus open-economy model presented in Gilles Oudiz and Jeffrey Sachs (1985), which is based on the Rudiger Dornbusch (1976) extension to the Mundell-Fleming model.³ If the subjective priors of policymakers (and of private agents) equal the objective probabilities, the average of welfare values achieved *ex post*—presumably the relevant criterion—will equal the *ex ante* expected welfare value. This was the case in our earlier paper; here, we relax that assumption and allow subjective priors and objective probabilities to differ, as do Frankel and Rockett (but we do not allow for disagreement concerning these probabilities, as they do). We estimate alternative versions of the Oudiz-Sachs model and perform stochastic simulations on the assumptions that one or the other version is the true one and that agents assign nonzero

probabilities to each of the models; we use the *ex post* welfare criterion in evaluating the gains from coordination. However, policymakers are assumed to take account of the model uncertainty and to maximize expected utility over the range of models.

This paper also makes a significant departure from all of the earlier coordination literature in that it abandons the purely static view of model uncertainty which has hitherto been adopted. Instead, it is assumed that agents update their priors over the set of possible models in a Bayesian fashion.⁴ There is an extensive literature on "learning rational expectations" (William Brock, 1972; R. M. Cyert and M. E. De Groot, 1974; Stephen De Canio, 1979; Lawrence Blume and David Easley, 1982; Blume et al., 1982; Margaret Bray, 1982; Bray and Neil Savin, 1986; Albert Marcet and Thomas Sargent, 1988). The conclusion of this literature is that least-squares learning seems to converge to rational expectations in a wide variety of circumstances but that it may also converge to an incorrect model. In particular, the paper by Blume and Easley is closest to the setup of our paper, as the authors consider Bayesian updating of prior probabilities applied to a finite set of models. They construct examples of cases in which both the true model and the wrong model are locally stable (i.e., cases in which it is possible that agents do not converge to rational expectations). It is thus of interest to analyze in an empirical macroeconomic model whether convergence to rational expectations occurs both in a cooperative and in a Nash equilibrium. Furthermore, since the process of learning affects both policy-setting and the private sector's expectation formation, it will change

²In a section entitled, "Extensions with Uncertainty," Frankel and Rockett (1988) also consider cases in which models are not believed to be correct with certainty; however, priors are not rational in that the true objective probability does not equal the subjective prior.

³The multicountry models surveyed by Frankel (1988) are generally elaborations of this simple model structure.

⁴Frankel and Rockett (1988 p. 318) refer to this possibility in justifying their assumption of disagreement among policymakers but do not treat it formally: "If one wishes to think of actors as perpetually processing new information in a Bayesian manner, so that their models over time would converge on any given reality in the limit, then one must admit that the speed of convergence is sufficiently slow, or else that reality is changing sufficiently rapidly, that policymakers have not been able to reach agreement on the true model."

the economy's equilibrium and hence the estimated gains from coordination.

We evaluate three regimes: a cooperative equilibrium in which monetary policies (in particular, the monetary base) are jointly chosen to maximize a weighted average of the two countries' utilities; a noncooperative, or Nash, equilibrium in which each country maximizes its own utility, taking as given the actions of the foreign government; and finally a noninterventionist "pure float" exchange-rate policy, in which each country keeps the money supply at its exogenous target level. The models we use are based on an empirically estimated, two-country (United States and an aggregate rest-of-OECD [Organization for Economic Coordination and Development]) model with a number of structural variants. Although these structural differences are seemingly minor and innocuous, the differences in the reduced-form multipliers of the models are substantial, with the degree of model uncertainty similar to that in Frankel (1988).⁵ Policy conflicts between the countries arise from structural shocks to the world economy, including shocks to each country's money demand, aggregate demand, and inflation and to their joint exchange rate. In our simulation analysis, we use drawings from a joint distribution describing these shocks that is based on the empirically estimated covariance matrix.

Our main conclusions may be briefly summarized. First, if optimal policies—whether cooperative or noncooperative—assign sufficiently little weight to the true model and there is no learning, then the economy can become dynamically unstable with potentially large gains or losses from coordination. In contrast, uncoordinated policies involving freely floating exchange rates and exogenous money targets never result in dynamic instabilities in our estimated version of the Oudiz-Sachs model. Therefore, in the absence of model learning, such a

policy regime may be optimal, because it is more robust to model misspecification. We have argued previously (Ghosh and Masson, 1988) that the lesson from model uncertainty is not that coordination is bad, but that policy-setting—whether it is coordinated or not—should in some sense be more cautious. Our results here support that conclusion. Second, once Bayesian learning is allowed, optimal policies never result in dynamic instability, and even when very little initial probability is assigned to the true model, the *ex post* gains from coordination (when discounted back to the present) are *always* positive in our simulation exercises.

The reason for the latter result is simple. If policies are set to maximize *ex ante* expected utility, coordination only results in welfare deterioration *ex post* if the models are very different; in that case, however, it becomes very easy to distinguish between the models and to learn which is the true model. The subjective priors therefore quickly converge to the true model, and coordination is welfare-improving. We would not want to exaggerate the relevance of this result to real-world policy choice; instead, it highlights the inadequacy of the assumption that one model is the "true" one.⁶ Nevertheless, experience of the past quarter century does provide evidence that policymakers abandon views of the world that can be seen to be wrong (such as the view that there exists a long-run trade-off between unemployment and inflation that can be exploited by aggregate demand policy) and thereby avoid the more disastrous consequences of their actions.

I. The Empirical Model

The empirical model we adopt is a general, two-country Mundell-Fleming model with forward-looking exchange-rate expectations (see Dornbusch, 1976; Oudiz and

⁵The degree of model uncertainty for a particular reduced-form multiplier is measured in terms of the ratio of its squared average across models to the sum of its squared average and its variance.

⁶In the context of the quote in footnote 4, we would argue that the speed of learning is unlikely to be the source of the problem, but rather, that reality is much more complex than the models and is changing too rapidly.

Sachs, 1985). The model was estimated using the data that are the basis for the IMF's MULTIMOD model (Masson et al., 1988). Results are presented in Table 1, using the same notation as in Oudiz and Sachs (1985). Our two "countries" are the United States and the rest of the world (ROW); data for the latter region resulted from aggregation of the remaining industrial countries in MULTIMOD.⁷ The estimation period, using annual data, was 1966–1986. Aggregate demand equations were estimated using instrumental variables: the instruments included money stocks, government spending, time, and lagged prices and output.

The coefficient estimates are all of the right signs and are generally fairly well determined—in particular, the effects of the real exchange rate on U.S. aggregate demand and of U.S. GDP on rest-of-world demand, the money-demand parameters, and the change in GDP effects on the two regions' output price changes. Despite its simplicity, the model seems to fit the data fairly well. Nevertheless, residual serial correlation is evident in the money-demand equations, and the specification embodies arbitrary constraints that are open to discussion. We proceed to relax some of these restrictions below and to estimate the resulting alternative models.

The computational burden of calculating optimal policies with model learning severely restricted the number of alternative models we could introduce; in this paper, we consider three possible models. The two alternative models, models II and III, differ from the baseline model (model I) in two respects: in model II money balances are deflated by the consumer price index rather than the GDP deflator, and there is a lagged dependent variable in the money-demand

equation, while in model III there is also a nonvertical Phillips curve. Estimates for these models are also presented in Table 1.

A first arbitrary feature of model I relates to the proper deflator for real money balances. As William Branson and Willem Buiter (1983) point out in their discussion of the Mundell-Fleming model, the effects of monetary and fiscal policies can be importantly different depending on whether the domestic output price or a broader index that includes foreign goods is used. The Oudiz-Sachs specification, excluding as it does the effects of terms of trade, conforms in this respect to the original Mundell-Fleming model. Model II deflates money balances not by p but by p^c . A second major area of arbitrariness in the Oudiz-Sachs model is the dynamic specification. The money-demand equation, whether specified with the output or consumption deflator, shows evidence of residual serial correlation. Most studies have allowed for the possibility that money balances adjust with a lag and have included a lagged dependent variable (see e.g., Stephen Goldfeld, 1973); this is also done in model II. The resulting money-demand equations fit considerably better than the ones in model I.

Model III includes the changes of model II and also relaxes the unitary coefficient on the lagged rate of inflation in the domestic-price-change equation. Relaxing this restriction lowers the standard errors of estimate in both the U.S. and rest-of-world equations, and the standard errors for the inflation coefficients imply rejection of the unitary coefficient for both the United States and the rest of the world. Furthermore, the change in output becomes insignificant, so we dropped this variable. Though the fit of the equations is only marginally superior to that for model I, the two specifications for the inflation equation have quite different long-run properties. The Oudiz-Sachs model exhibits no long-run trade-off between output and inflation, as in the steady state both output prices and consumer prices grow at the same, steady rate. In contrast, with a nonunitary coefficient, different rates of inflation are associated with different rates of

⁷In general, variables were aggregated by converting to a common currency and summing. GDP (gross domestic product) weights were used to aggregate interest rates. The exchange rate was taken to be the reciprocal of the MERM-weighted effective rate (Jacques Artus and Anne McGuirk, 1981) of the U.S. dollar (in index form, 1980 = 1). The rest-of-world price level and money supply are expressed in this "currency."

TABLE 1—PARAMETER ESTIMATES FOR OUDIZ-SACHS MODEL AND VARIANTS,
1966–1986 (STANDARD ERRORS IN PARENTHESES)

Model I	
<i>Aggregate demand:</i>	
$q_t = -0.114 (p_t - e_t - p_t^*) + 0.054 q_t^* - 0.152 (i_t - p_{t+1} + p_t) + 0.020 t + 0.300 g_t + 5.196$ (0.050) (0.209) (0.340) (0.009) (0.166) (2.298)	$[\bar{R}^2 = 0.988, \sigma = 0.021, DW = 1.78]$
$q_t^* = 0.167 (p_t - e_t - p_t^*) + 0.550 q_t - 0.378 (i_t^* - p_{t+1}^* + p_t^*) - 0.031 t + 1.614 g_t^* - 6.077$ (0.099) (0.499) (0.212) (0.026) (0.477) (6.129)	$[\bar{R}^2 = 0.992, \sigma = 0.021, DW = 1.16]$
<i>Money demand:</i>	
$m_t - p_t = 0.225 q_t - 1.419 i_t + 3.453$ (0.066) (0.413) (0.503)	$[\bar{R}^2 = 0.382, \sigma = 0.038, DW = 1.02]$
$m_t^* - p_t^* = 0.700 q_t^* - 1.077 i_t^* + 0.100$ (0.078) (0.609) (0.616)	$[\bar{R}^2 = 0.853, \sigma = 0.051, DW = 0.67]$
<i>Consumer price index:</i>	
$p_t^c = 0.899 p_t + (1 - 0.899)(p_t^* + e_t)$	
$p_t^{c*} = 0.758 p_t^* + (1 - 0.758)(p_t - e_t)$	
<i>Output price change:</i>	
$p_t - p_{t-1} = (p_{t-1}^c - p_{t-2}^c) + 0.095 [\hat{q}_{t-1} - 0.027(t-1) - 7.373] + 0.309 (q_{t-1} - q_{t-2}) - 0.009$ (0.119) (0.108) (0.004)	$[\bar{R}^2 = 0.351, \sigma = 0.011, DW = 2.33]$
$p_t^* - p_{t-1}^* = (p_{t-1}^{c*} - p_{t-2}^{c*}) + 0.040 [q_{t-1}^* - 0.033(t-1) - 7.803] + 0.880 (q_{t-1}^* - q_{t-2}^*) - 0.031$ (0.088) (0.204) (0.008)	$[\bar{R}^2 = 0.458, \sigma = 0.017, DW = 1.74]$
<i>Exchange rate:</i>	
$e_{t+1} = e_t + i_t - i_t^* + \varepsilon$	$[\sigma_\varepsilon = 0.026]$
Model II	
(All equations as for model I except the following)	
<i>Money demand:</i>	
$m_t - p_t^c = 0.184 q_t - 1.387 i_t + 0.680 (m_{t-1} - p_{t-1}^c) + 0.302$ (0.051) (0.318) (0.160) (0.857)	$[\bar{R}^2 = 0.691, \sigma = 0.029, DW = 1.63]$
$m_t^* - p_t^{c*} = 0.277 q_t^* - 1.579 i_t^* + 0.702 (m_{t-1}^* - p_{t-1}^{c*}) - 0.421$ (0.094) (0.456) (0.141) (0.494)	$[\bar{R}^2 = 0.909, \sigma = 0.038, DW = 1.65]$
Model III	
(All equations as for model II except the following)	
<i>Output price change:</i>	
$p_t - p_{t-1} = 0.611 (p_{t-1}^c - p_{t-2}^c) + 0.364 [\hat{q}_{t-1} - 0.027(t-1) - 7.373] + 0.022$ (0.098) (0.101) (0.006)	$[\bar{R}^2 = 0.783, \sigma = 0.010, DW = 1.95]$
$p_t^* - p_{t-1}^* = 0.238 (p_{t-1}^{c*} - p_{t-2}^{c*}) + 0.491 [q_{t-1}^* - 0.033(t-1) - 7.803] + 0.051$ (0.146) (0.116) (0.010)	$[\bar{R}^2 = 0.761, \sigma = 0.015, DW = 1.44]$
Variable Definitions	
(U.S. variables are unstarred; non-U.S. variables are starred; all variables except i and t are in logs)	
e = nominal exchange rate (dollars per foreign currency)	
g = real government spending on goods and services	
i = nominal short-term interest rate	
m = monetary base	
p = GDP deflator	
p^c = deflator of domestic absorption	
q = real GDP	
t = time trend	

TABLE 2—EIGENVALUES OF MODELS I–III

Model I	Model II	Model III
1.421	1.237	1.195
0.975	0.982 ± 0.069i	0.942
0.936 ± 0.119i	0.974	0.782
−0.085	0.611	0.656
	−0.363	0.237
	−0.044	

output growth; this possibility would be disputed by many economists, however (see e.g., Milton Friedman, 1968). On the basis of the estimation results, it may be reasonable to attribute some nonzero probability to the existence of such a trade-off.

Although the structural differences among the three models seem minor, the reduced-form multipliers differ considerably across the models, as do the dynamic properties (see Table 2 for the eigenvalues of the three models). There is one unstable eigenvalue in each case, corresponding to the one nonpredetermined variable, the exchange rate. Model II has a more complicated dynamic structure and two extra real eigenvalues, compared to model I: this is due to lags in money demand. Model III retains the money-demand equation but has a simpler adjustment process for prices.

A simple summary statistic for the degree of model uncertainty is given by the ratio

$$(1) \quad \zeta = \mu^2 / (\mu^2 + \sigma^2)$$

where μ is the mean multiplier of an instrument on a particular target and σ^2 is its variance (across the three models). This gives a dimensionless statistic which equals unity if there is no model uncertainty and approaches zero as the effectiveness of the policy instrument deteriorates, in the sense of William Brainard (1967). In a model with forward-looking variables, the multipliers depend upon the anticipation of future policies as well as current policies and therefore are not independent of the regime under consideration (it is assumed for this purpose that a policy of exogenously setting the money supply prevails).

Table 3 includes the second-year multipliers for the three models, as well as ζ values, for an exogenous permanent increase in the money supply of 4 percent. As expected from theoretical models, the transmission effects of a monetary expansion can have either sign, depending on parameter values: models I and II imply positive transmission onto foreign output, while the reverse is true of model III. Somewhat surprisingly, the second-year *domestic* output effects of a ROW money increase also are ambiguous: the negative effect on ROW output in models I and II results from the large coefficient on the lagged change in output in the ROW inflation equation, which pushes up domestic output prices by enough that the real exchange rate appreciates in the second year relative to baseline.⁸ In model III, ROW output effects of a ROW monetary expansion are persistently positive. Though models I and II give the same signs for all the multipliers, their numerical values differ substantially.

It is interesting to compare the ζ values for our three models with those implied by the models in Frankel's (1988) study, also calculated for such a money-supply shock. As can be seen from Table 3, his study of 12 different models incorporates a somewhat greater degree of uncertainty about the reduced-form multipliers of a U.S. monetary expansion than our three alternative models, but the output effects of a ROW monetary expansion are more certain in Frankel's set of models than in ours.

II. Optimal Policies Under Model Uncertainty and Model Learning

In order to calculate the average *ex post* gains from coordination with model learning, we adapt the algorithm we developed in Ghosh and Masson (1988). The logic of the model is as follows. In period t , the state vector \mathbf{x}_t and a vector of subjective priors

⁸In the *first* year, the money supply increase leads to a large increase in ROW output, however: by 2.8 percent in model I and by 2.2 percent in model II.

TABLE 3—REDUCED-FORM MODEL MULTIPLIERS AND MEASURES OF MODEL UNCERTAINTY

Model	Effects of U.S. money supply on:				Effects of ROW money supply on:			
	U.S. Y	ROW Y	U.S. P	ROW P	U.S. Y	ROW Y	U.S. P	ROW P
<i>Second-year effects of a 4-percent monetary expansion (percentage):</i>								
I	0.5	1.0	1.2	-2.6	0.2	-0.1	-0.8	3.8
II	0.2	0.2	0.9	-1.8	0.1	-0.4	-0.6	3.1
III	0.5	-0.2	0.8	-0.4	-0.2	0.5	-0.5	1.0
<i>Measures of model uncertainty (ζ values; percentage):^a</i>								
I-III	88	31	97	75	4	— ^b	96	83
Frankel (1988)	68	— ^b	45	52	19	72	24	25

^aBased on second-year effects; a smaller number indicates greater uncertainty.

^bThe effects were clustered around zero, giving a zero value for ζ .

Π_t (with elements π_t^i) are inherited. Policymakers choose a vector of controls (i.e., policy instruments) u_t in order to influence their target vector τ_t . They do not attempt active learning (i.e., performing policy experiments in order to discover which of the models is correct). At the end of period t , a vector of endogenous variables ω_{t+1} is observed which allows agents to update their priors, yielding Π_{t+1} .

The dynamics of the world model are assumed to be given by:

$$(2) \begin{bmatrix} x_{t+1}^i \\ e_{t+1}^i \end{bmatrix} = \begin{bmatrix} A^i & B^i \\ D^i & F^i \end{bmatrix} \begin{bmatrix} x_t^i \\ e_t^i \end{bmatrix} + \begin{bmatrix} C^i \\ G^i \end{bmatrix} u_t + \begin{bmatrix} \Theta_x^i \\ \Theta_e^i \end{bmatrix} \varepsilon_t$$

for i ranging over the possible models i, \dots, k , and where x is a vector of state variables, e is a vector of jumping variables, u is the vector of controls, and ε is an unobserved vector white-noise shock, distributed $N(0, \Sigma)$ (Σ is the variance-covariance matrix). $A^i, B^i, C^i, D^i, F^i, G^i, \Theta_x^i$, and Θ_e^i are constant matrices associated with model i .

In addition, structural equations of the models map the state variables and the forward-looking variables into a vector of targets τ :

$$(3) \tau_t^i = L^i x_t + M^i e_t + N^i u_t + \Theta_\tau^i \varepsilon_t.$$

Policymakers in each country are assumed to have preferences over the target vector which are represented by

$$(4) v = \max \left[- (1/2) \sum_{t=0}^{\infty} \beta^t E \{ \tau_t' \Omega \tau_t \} \right]$$

and

$$(5) v^* = \max \left[- (1/2) \sum_{t=0}^{\infty} \beta^t E \{ \tau_t' \Omega^* \tau_t \} \right]$$

respectively, where β is a discount factor, Ω and Ω^* are matrices that weight the target variables, and the expectation is taken with respect to uncertainty about both the correct model and the current realization of the shock ε_t . We assume that, although agents do not know the true model of the world economy, they do know the variance-covariance matrix of the additive shocks.⁹ Moreover, we do not allow heterogeneity of agents; both the private sectors and the governments start with the same priors across models and update them in the same fashion.

Following Oudiz and Sachs (1985), we derive the optimal linear decision rules by first calculating the dynamic programming

⁹This is a somewhat heroic assumption, but it simplifies the analysis considerably. Furthermore, our main conclusions do not appear to be excessively sensitive to the assumed variance-covariance matrix.

solution¹⁰ for a (finite) T -period horizon and then increasing T until stationary matrices Γ and Γ^* (functions of Π_t) are obtained:

$$(6) \quad \mathbf{u}_t = \Gamma(\Pi_t)\mathbf{x}_t \quad \text{and} \quad \mathbf{u}_t^* = \Gamma^*(\Pi_t)\mathbf{x}_t.$$

Moreover, this allows us to write the vector of forward-looking variables as a linear function of a matrix λ and the state vector \mathbf{x}_t :

$$(7) \quad \mathbf{e}_t = \lambda(\Pi_t)\mathbf{x}_t.$$

From equations (2)–(7), we can then express the value functions as quadratic forms in \mathbf{x}_t and matrices \mathbf{S} and \mathbf{S}^* (dependence on \mathbf{x}_t and Π_t is made explicit by the functional notation):

$$(8) \quad v(\mathbf{x}_t, \Pi_t) = \mathbf{x}_t' \mathbf{S}(\Pi_t) \mathbf{x}_t$$

$$v^*(\mathbf{x}_t, \Pi_t) = \mathbf{x}_t' \mathbf{S}^*(\Pi_t) \mathbf{x}_t.$$

Once the optimal stationary policy rules have been obtained, the model is simulated forward, and priors are updated using Bayesian inference. The forward simulation is conditional on a particular model, say model j , being true. Suppose that in period t , the world economy has inherited the state \mathbf{x}_t and priors over the models are given by Π_t . The optimal policy in period t for the home country is given by

$$(9) \quad \mathbf{u}_t = \Gamma(\Pi_t)\mathbf{x}_t.$$

A drawing from the shocks \mathbf{e}_t is made, and the state vector in $t+1$ is therefore given by

$$(10) \quad \mathbf{x}_{t+1} = \mathbf{A}^j \mathbf{x}_t + \mathbf{B}^j \lambda^j(\Pi_t) \mathbf{x}_t \\ + \mathbf{C}^j \mathbf{u}_t + \Theta_{\mathbf{x}}^j \mathbf{e}_t.$$

At the beginning of period $t+1$, agents

observe a vector of variables ω_{t+1}^i . Each of the k possible models implies a structural relationship for the observation vector ω_{t+1}^i

$$(11) \quad \omega_{t+1}^i = \mathbf{W}_{\mathbf{x}}^i \mathbf{x}_t + \mathbf{W}_{\mathbf{e}}^i \mathbf{e}_t + \mathbf{W}_{\mathbf{u}}^i \mathbf{u}_t + \mathbf{W}_{\mathbf{e}}^i \mathbf{e}_t,$$

where matrices $\mathbf{W}_{\mathbf{x}}^i$, $\mathbf{W}_{\mathbf{e}}^i$, $\mathbf{W}_{\mathbf{u}}^i$, and $\mathbf{W}_{\mathbf{e}}^i$ are obtained from the relevant rows of (2) and it is assumed that $\mathbf{W}_{\mathbf{e}}^i$ is invertible.¹¹

Let $E(\omega_{t+1}^i)$ be the expected value of ω_{t+1}^i [evaluated at $E(\mathbf{e}_t) = \mathbf{0}$]. The value of the shock implied by each model is therefore

$$(12) \quad \mathbf{e}_t^i = (\mathbf{W}_{\mathbf{e}}^i)^{-1} [\omega_{t+1}^i - E(\omega_{t+1}^i)].$$

The new Bayesian priors are then given by

$$(13) \quad \pi_{t+1}^i = \frac{\Pr(\mathbf{e}_t^i | \Sigma^i) \pi_t^i}{\sum_{i=1}^k \Pr(\mathbf{e}_t^i | \Sigma^i) \pi_t^i}$$

where $\Pr(\cdot)$ is the probability that a vector shock, distributed $N(0, \Sigma^i)$, takes the value \mathbf{e}_t^i . The state variables in period $t+2$ are then generated with a drawing for \mathbf{e}_{t+1} , and the whole process is repeated.¹²

III. Simulation Results

In addition to the model parameters, the simulation analysis requires specification of the policymakers' discount factors (chosen to be 0.95; i.e., a discount rate of 5 percent per annum), the utility weights on each target, and the relative weight each country receives in the social planner's objective function. The utility weights were taken from the revealed preference estimates of Oudiz and Sachs (1984) with policymakers assumed to target inflation and output.¹³

¹¹If agents observe fewer variables than the number of shocks in \mathbf{e}_t , then they face a signal-extraction problem concerning the shocks as well as the models.

¹²Until the priors converge to the model generating the data, expectations will not be unbiased.

¹³The current-account objective in Oudiz and Sachs (1984) was ignored. The utility-function weights used were those reported in table 11 of their paper; for the rest of the world, weights for Japan and West Germany were averaged.

¹⁰The dynamic programming solution excludes trigger mechanisms and reputational strategies. It may therefore exaggerate the gains from coordination relative to Nash solutions (see Matthew Canzoneri and Dale Henderson, 1988).

The objective functions are as follows for the United States (unstarred) and the rest of the world (starred):¹⁴

(14)

$$V = E \left\{ \sum_{t=0}^{\infty} 0.95^t [0.07q_t^2 + 0.49(\Delta p_t^c)^2] \right\}$$

(15)

$$V^* = E \left\{ \sum_{t=0}^{\infty} 0.95^t [0.045q_t^{*2} + 0.50(\Delta p_t^{c*})^2] \right\}.$$

The world objective function is

$$(16) \quad V_w = \alpha V + (1 - \alpha)V^*.$$

In simulations reported below, $\alpha = 0.5$; each country is given equal weight in the world objective function, which is used to evaluate the outcomes of the three policy regimes.¹⁵

In the Nash equilibrium, the United States maximizes (14) with respect to m_t , and the rest of the world maximizes (15) with respect to m_t^* . In the cooperative regime, (16) is maximized with respect to $\{m_t, m_t^*\}$. In the floating regime, policy is

described by an exogenous money-supply target, whatever the value of the objective functions. The observation vector includes output, interest rates, and producer prices in each country and the nominal exchange rate. The optimal policies and value functions depend nonlinearly on the probabilities assigned to each of the three models. The recursive optimization had to be done on a two-dimensional grid; that is, for each possible combination $[\pi^1, \pi^2, (1 - \pi^1 - \pi^2)]$, the recursive algorithm outlined above must be solved.¹⁶

In the first set of simulations, all agents assign an initial probability to each model and do not update their priors (to repeat, we assume throughout that all agents in the model—both private and public—share a common set of subjective priors). The results are reported conditional on each of the three models being the true model; we also cite results for the certainty case, in which agents assign a probability of 1 to the true model. The values reported in Table 4 represent the total present value of disutility for the world economy (with equal weights on the United States and the rest of the world) and are expressed in terms of GDP equivalents. As is evident below, some of the simulations exhibit explosive behavior, and the present value of disutility may be undefined; in these cases, we have simply marked the entry “explosive.” Since there is an additive random shock, ϵ_t , disutility depends upon the specific realizations (which are drawn from a normal distribution with zero mean and the estimated variance-covariance matrix of the “true” model),¹⁷ and hence we have taken the average of ten stochastic simulations (the drawings are the same for each of the models). The optimal

¹⁴ Variables q and Δp^c are output and the rate of change of consumer prices, both as deviations from the baseline.

¹⁵ The welfare values in Table 4 are all for world welfare, using the same equally weighted objective function. There is no guarantee that both countries' welfares increase as a result of cooperation (although in most cases both do). We have verified that, in all cases where coordination improves world welfare, a set of weights could be found such that each country's welfare (averaged over the ten drawings of the shocks) was higher. In practice, when unequal weights were necessary, a greater weight had to be given to the United States than to the rest of the world in world welfare, no doubt a result of asymmetries in the multipliers reported in Table 3. Such welfare weights, however, lowered the overall gains from cooperation, when evaluated using a consistently defined objective function. Alternatively, if a mechanism for side payments existed, then both countries would always gain, even when cooperation involved maximizing the equally weighted welfare function.

¹⁶ We used a grid of 11 equally spaced intervals. Table 4 reports a subset of those simulations, plus simulations in which a very small weight (0.045) is placed on the true model, and others in which a weight of 0.5 is placed on the true model and weights of 0.25 are placed on the alternative models.

¹⁷ None of the estimated correlations between shocks was significant at the 5-percent level, so a diagonal variance-covariance matrix was used to generate the shocks.

TABLE 4—DISUTILITY LEVELS WITH AND WITHOUT MODEL-LEARNING

Probabilities assigned to each model ^a			Regime ^c		
I	II	III	Cooperative	Nash	Float
No model-learning:					
1.0*	0.0	0.0	31	42	44
0.5*	0.25	0.25	32	41	43
0.1*	0.9	0.0	39	45	43
0.1*	0.0	0.9	40	87	43
0.045*	0.0	0.955	explosive	explosive	47
0.0	1.0*	0.0	42	51	55
0.25	0.5*	0.25	58	71	55
0.9	0.1*	0.0	67	82	55
0.0	0.1*	0.9	55	explosive	55
0.0	0.045*	0.955	explosive	explosive	55
0.0	0.0	1.0*	21	28	24
0.25	0.25	0.5*	22	28	24
0.9	0.0	0.1*	23	31	25
0.0	0.9	0.1*	24	28	24
Model-learning: ^b					
0.5*	0.25	0.25	31	41	44
0.1*	0.9	0.0	31	41	44
0.1*	0.0	0.9	31	41	44
0.045*	0.0	0.955	31	41	44
0.25	0.5*	0.25	42	51	56
0.9	0.1*	0.0	42	52	56
0.0	0.1*	0.9	41	53	55
0.0	0.045*	0.955	42	51	55
0.25	0.25	0.5*	21	28	24
0.9	0.0	0.1*	21	29	24
0.0	0.9	0.1*	22	28	24

^aThe correct model is indicated by an asterisk.

^bTable entries are initial probabilities; probabilities are updated in a Bayesian fashion.

policies are designed to stabilize output and inflation against the specific random structural shocks applied to the model; as such, the optimal policies under both cooperative and noncooperative behavior are not easily interpretable and are therefore not reported.

There are two noteworthy points about the top part of Table 4 (no model learning). First, when equal weights are assigned to the competing models and a weight of 0.5 is assigned to the "true" one, the gains from coordination relative to the Nash equilibrium are not spectacular but are certainly measurably positive: they are comparable in size to those under certainty (the top line of each panel). These gains, which range from 6 percent to 13 percent of GDP, amount to a permanent increase of about half of 1

percent of GDP per year (using the 5-percent discount rate), an estimate which is in line with those of previous studies (e.g., Oudiz and Sachs [1984], who calculate welfare gains in that form to be less than 1 percent).¹⁸ Both the Nash and the cooperative equilibria are significantly better than a pure float regime for model I; this is not true of model III, however, where floating is intermediate between the other two regimes. For model II, under certainty, the Nash equilibrium is better than floating; but with a probability of 0.5 on model II and 0.25 on models I and II, the Nash equilibrium is

¹⁸In Oudiz and Sachs (1984), the gains are given in terms of GDP equivalents sustained for three years.

considerably worse and the cooperative regime is somewhat worse than the pure float regime.

Second, when we assign a low probability to the true model and a high probability to one of the competing models, the cooperative regime actually yields substantially lower *ex post* welfare than the pure float regime. The situation is most critical when a large weight is attached to model III (and either model I or model II is correct); here, the cooperative solution results in (eventually) explosive behavior¹⁹ of the economy or in large disutility. The Nash equilibrium also implies large (and in some cases, unbounded) disutility when the weights are over 90 percent on model III but one of the other models is correct. In fact, the most robust policy to follow is one of complete nonintervention. The large losses and possible instability no doubt result from the fact that model III implies a long-run trade-off between output and inflation, which activist policies (either Nash or cooperative) try to exploit—unsuccessfully, it turns out, because the true model, either model I or model II, is in fact a natural-rate model. In either case, coordination severely reduces welfare to the extent of making the sustainability of a coordinated regime infeasible. However, it is important to note that it is not only the cooperative regime that may result in low welfare when policymakers use the wrong model. For example, if model II is the true model but a high prior (0.9) is put on model III, the Nash equilibrium is highly inefficient, so that there are very large *gains* from coordination. In the cases when both cooperative and Nash equilibria exhibited instability, the disutility levels were larger and diverged more quickly for the latter.

The conclusion that emerges from the above results is therefore that a policy rule that does not require active intervention, corresponding to exogenous money supplies

and floating exchange rates, is the safest policy when policymakers are uncertain about the model. Our results therefore provide some support to the advocacy of fixed rules in preference to activist policies, which Friedman has long argued may be destabilizing (see e.g., Friedman, 1948).²⁰

The results suggest that, when policymakers have the wrong priors over the models and do not undertake updating, any policy intervention—coordinated or uncoordinated—can be dangerous. However, it is implausible that policymakers would not update their subjective priors when they find that their expectations about the effects of their policies are consistently invalidated. In a second set of simulations, therefore, we assume that agents update their priors over the models in the Bayesian fashion described above. The lower part of Table 4 reports the discounted present value of disutility under cooperative and noncooperative behavior, as well as the floating-exchange-rate regime with exogenous money targets, for simulations with various initial probability priors.

The important conclusion that emerges is that policies no longer become unstable, even when the initial priors attributed to the true model are very low. Furthermore, we find that coordination is always welfare-improving relative to both the Nash equilibrium and nonactivist policies. Consider, for example, the case in which agents assign a prior probability of 4.5 percent to model I and 95.5 percent to model III, when model I is the correct model. In the absence of Bayesian updating, coordination reduces welfare substantially relative to a pure float, since the coordinated regime is dynamically unstable. Bayesian learning, combined with the same initial priors, results in a welfare gain relative to both Nash equilibrium and floating.²¹ With endogenous model-learning,

¹⁹The instability manifested itself in a failure of policy settings to converge, and in the fact that disutility levels were not bounded as the horizon was extended.

²⁰There are other considerations that are relevant to the choice between fixed rules and more activist policies (see e.g., Stanley Fischer, 1988).

²¹Nevertheless, a superior policy *ex post* would have been to follow a noninterventionist policy until policymakers reduced their probability weight on the incor-

ing, the discounted-present-value gains from coordination are all in the range of 7–15 percent of GDP.²² The noninterventionist-policy regime now performs consistently worse than the cooperative regime but is sometimes better than the Nash equilibrium—in particular, when model III (the non-vertical Phillips curve model) is true. In this case, it would seem that the scope for beggar-thy-neighbor policies is greater; they are ruled out by both the cooperative and “floating” (i.e., nonactivist) regimes.

The examples in which policymakers (and the private sector) assign a high initial weight to model III, when in reality one of the other two models is correct, have some relevance to the history of demand management in the postwar period. Early models of the “Phillips curve” (see e.g., A. W. Phillips, 1958; Richard Lipsey, 1960) implied that there was a trade-off between the rate of change of wages or prices and the unemployment rate or output. These models no doubt helped induce central banks and treasuries to engage in demand expansion, in an attempt to buy more output growth at what was judged to be an acceptable inflation cost. The experience of accelerating inflation beginning in the late 1960’s forced economists and policymakers to reconsider those models, and there has been a profound shift in policy away from short-term fine-tuning and to a concern for the medium-term inflation consequences of policy. Moreover, the rationale for the policy changes has been acceptance of natural-rate models, which do not allow for monetary stimulus to have permanent positive effects on the level of activity. Friedman (1977 p. 470) commented on the change in policy in the following terms in his Nobel lecture:

Government policy about inflation and unemployment has been at the center

of political controversy. Ideological war has raged over these matters. Yet the drastic change that has occurred in economic theory has not been a result of ideological warfare. It has not resulted from divergent political beliefs or aims. It has responded almost entirely to the force of events: brute experience proved far more potent than the strongest of political or ideological preferences.

IV. Caveats and Discussion

Having found that, in the presence of model-learning, the cooperative regime dominates noncooperation and nonintervention in the context of our model, we tried to gauge the sensitivity of the results to our assumptions. The *ex post* performance of the coordinated regime depends on two factors: the robustness of the optimal policy to model errors; and the rate of learning of the model. Clearly, the more diverse the models are, the more likely the optimal policy based on the “wrong” priors is to result in lower welfare. We also thought that we might have assumed too little uncertainty in choosing the structurally similar models of Section III. However, in further simulations with more diverse models, we found that as the degree of model uncertainty increased (in the sense of lower ζ ratios) the rate of model-learning also increased. The intuition is straightforward: if the models are very different, then the implied observation vector of each model is also very different, and the updated priors assigned to the false models will be correspondingly low. Thus, although greater diversity between the models makes more likely the possibility of welfare-deteriorating policies, it also serves to reduce the model uncertainty (at least when there is a finite number of alternative models, as is the case here).

It is possible to decrease the rate of learning by increasing the variance of the additive shocks ϵ_t . The greater the variance of ϵ_t , the greater the noise in the updating observations and, therefore, the slower the rate of learning. We checked the sensitivity of our results by using a variance matrix with ten times the estimated standard er-

rect model (model III) sufficiently and then cooperated.

²² When the world objective function uses weights that ensure that welfare averaged over the ten stochastic simulations improves in both countries as a result of cooperation, the range of gains is 1.5–15 percent.

rors. Despite this large increase in error variances, however, we were unable to reverse our conclusions about the ranking of the coordinated and uncoordinated regimes in the presence of learning.

It must be emphasized, however, that our experiments assume knowledge of the variance-covariance matrix describing the shocks. Forcing agents to estimate it would introduce considerably more uncertainty and make it more difficult to infer which model was correct, as would a combination of temporary and permanent shocks.

Another caveat is that we have simplified the problem to one in which policymakers must just discover the unique unchanging model describing the economy. An alternative assumption is that agents can never perfectly anticipate the true model when setting policies, because the true model is stochastic. If the vector of subjective priors converges to the true probability that the model is realized, then the average *ex post* welfare gain will equal the *ex ante* expected welfare gain from coordination, at least in large samples. In that case, we return to the *ex ante* expected welfare criterion in which coordination is necessarily welfare-improving.²³ More realistic is the case in which the true model is changing in a nonrandom way as a result of structural shifts and in which these shifts occur frequently enough so that the distribution describing the models is never completely learned. We have yet to explore this case.

V. Conclusions

In this paper, we have discussed whether coordination is likely to reduce the actual *ex post* level of welfare when policymakers are uncertain about the effects of policies. We have found no evidence in our simulations that policy coordination is likely to reduce welfare vis-à-vis noncooperative policies; however, a simple nonintervention

regime, such as a pure floating exchange rate accompanied by exogenous money targets, may be the most robust policy in the presence of model uncertainty. These conclusions are of course specific to a particular model and are based here on an estimated version of the Mundell-Fleming model with sticky prices and rational exchange-rate expectations. More experimentation with other models is no doubt necessary in order to gauge whether the conclusions can be generalized.

Once we introduced endogenous model-learning, we found that coordination always results in higher welfare: though the gains from coordination are not spectacular, they appear to be significantly positive. We were unable to generate losses from coordination, even by increasing the variance of the additive noise. No doubt, the representation of the economy is much too simple. What the results suggest, however, is that the conclusion that coordination is as likely to decrease welfare as to increase it (see Frankel and Rockett, 1988) is not consistent with the joint assumptions that there is a single "true" model and that agents learn about that model in a Bayesian fashion.

REFERENCES

- Artus, Jacques and McGuirk, Anne, "A Revised Version of the Multilateral Exchange Rate Model," *Staff Papers* (International Monetary Fund), June 1981, 28, 275-309.
- Blume, Lawrence E. and Easley, David, "Learning to Be Rational," *Journal of Economic Theory*, April 1982, 26, 340-51.
- _____, Bray, Margaret and Easley, David, "Introduction to the Stability of Rational Expectations. Equilibrium," *Journal of Economic Theory*, April 1982, 26, 313-7.
- Brainard, William C., "Uncertainty and the Effectiveness of Policy," *American Economic Review*, May 1967, 57, 411-25.
- Branson, William H. and Buiter, Willem H., "Monetary and Fiscal Policy with Flexible Exchange Rates," in J. Bhandari and B. Putnam, eds., *Economic Interdependence and Flexible Exchange Rates*, Cambridge,

²³This was verified in simulations in which the true model was drawn stochastically from the three possible models, each of which had equal probability, and agents undertook Bayesian learning as above.

- MA: MIT Press, 1983, pp. 251–85.
- Bray, Margaret, "Learning, Estimation, and the Stability of Rational Expectations," *Journal of Economic Theory*, April 1982, 26, 318–39.
- and Savin, Neil E., "Rational Expectations Equilibria, Learning, and Model Specification," *Econometrica*, September 1986, 54, 1129–60.
- Brock, William A., "On Models of Expectations That Arise from Maximizing Behavior of Economic Agents Over Time," *Journal of Economic Theory*, December 1972, 5, 348–76.
- Canzoneri, Matthew B. and Henderson, Dale W., "Is Sovereign Policymaking Bad?" in K. Brunner and A. H. Meltzer, eds., *Stabilization Policies and Labor Markets*, Carnegie-Rochester Conference Series on Public Policy, New York: North Holland, 1988, pp. 93–140.
- Cyert, R. M. and De Groot, M. E., "Rational Expectations and Bayesian Analysis," *Journal of Political Economy*, May/June 1974, 82, 521–36.
- De Canio, Stephen J., "Rational Expectations and Learning from Experience," *Quarterly Journal of Economics*, February 1979, 93, 47–57.
- Dornbusch, Rudiger, "Expectations and Exchange Rate Dynamics," *Journal of Political Economy*, December 1976, 84, 1161–76.
- Feldstein, Martin, "Distinguished Lecture on Economics in Government: Thinking about International Economic Coordination," *Journal of Economic Perspectives*, Spring 1988, 2, 3–13.
- Fischer, Stanley, "Rules versus Discretion in Monetary Policy," NBER (Cambridge, MA) Working Paper No. 2518, February 1988.
- Frankel, Jeffrey, "Ambiguous Policy Multipliers in Theory and in Empirical Models," in Ralph Bryant, Dale Henderson, Gerald Holtham, Peter Hooper, and Steven Symansky, eds., *Empirical Macroeconomics for Interdependent Economies*, Washington, DC: The Brookings Institution, 1988, pp. 17–26.
- , and Rockett, Katharine, "International Macroeconomic Policy Coordination when Policymakers Do Not Agree on the True Model," *American Economic Review*, June 1988, 78, 318–40.
- Friedman, Milton, "A Monetary and Fiscal Framework for Economic Stability," *American Economic Review*, June 1948, 38, 245–64; reprinted in Milton Friedman, *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953, pp. 133–56.
- , "The Role of Monetary Policy," *American Economic Review*, March 1968, 58, 1–17.
- , "Inflation and Unemployment," *Journal of Political Economy*, June 1977, 85, 451–72.
- Ghosh, Atish and Masson, Paul R., "International Policy Coordination in a World with Model Uncertainty," *Staff Papers* (International Monetary Fund), June 1988, 35, 230–58.
- and Ghosh, Swati, "Does Model Uncertainty Really Preclude International Policy Coordination?" *Journal of International Economics*, 1991, forthcoming.
- Goldfeld, Stephen, "The Demand for Money Revisited," *Brookings Papers on Economic Activity*, 1973, (3), 577–637.
- Holtham, Gerald and Hughes Hallett, Andrew, "International Policy Coordination and Model Uncertainty," in Ralph Bryant and Richard Portes, eds., *Global Macroeconomics: Policy Conflict and Cooperation*, London: Macmillan, 1987, pp. 128–77.
- Lipsey, Richard G., "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1862–1957: A Further Analysis," *Economica*, February 1960, 27, 1–31.
- Marcet, Albert and Sargent, Thomas J., "The Fate of Systems with 'Adaptive' Expectations," *American Economic Review*, May 1988, 78, 168–72.
- Masson, Paul R., Symansky, Steven, Haas, Richard and Dooley, Michael, "MULTIMOD: A Multi-Region Econometric Model," *Staff Studies for the World Economic Outlook* (Washington, DC: International Monetary Fund), July 1988, 50–104.
- Oudiz, Gilles and Sachs, Jeffrey, "Macroeconomic Policy Coordination Among Indus-

trial Economies," *Brookings Papers on Economic Activity*, 1984, (1), 1-75.

_____ and _____, "International Policy Coordination in Dynamic Macroeconomic Models," in Willem H. Buiter and Richard C. Marston, eds., *International Economic Policy Coordination*, Cam-

bridge: Cambridge University Press, 1985, pp. 274-319.

Phillips, A. W., "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861-1957," *Economica*, November 1958, 25, 283-99.

Welfare Dominance: An Application to Commodity Taxation

By SHLOMO YITZHAKI AND JOËL SLEMROD*

In this paper, we propose a method to identify commodity-tax changes that will be favored by all individuals who can agree on certain weak assumptions with regard to the social-welfare function. The method is based on an extension of the criterion of second-degree stochastic dominance and is illustrated using data from Israel. (JEL H21)

A principal weakness of the theory of optimal taxation with heterogeneous taxpayers is the dependence of the optimum tax rates on the exact properties of the social-welfare function. While other components of the problem, such as the excess burden of the tax system, can presumably be reconstructed from empirical observations on the behavior of consumers, it is clear that estimating the social-welfare function is not an easy task. Although there have been several attempts to recover the social-welfare function using the revealed preferences of governments (usually by assuming that governments act optimally according to a "just" principle of taxation, such as equal sacrifice), all these methods require very strong assumptions which are unlikely to command wide support (see Richard Musgrave, 1959; Koichi Mera, 1969; Burton Weisbrod, 1968; John Piggott, 1982; Menahem Yaari, 1988).

This problem is especially disturbing for developing countries that rely heavily on commodity taxes as a major policy instrument for raising revenues, dealing with poverty, and changing the income distribu-

tion. In most of these countries, data are not available even for estimating the excess burden of the tax system. Therefore, it seems that the theory of optimal commodity taxation is not very helpful as an input into policy formulation in this context. This problem is even more complicated from the perspective of an economic adviser whose role is to advise the government of a specific country. If, in an ideal case, the government can supply him or her with all the necessary data, then the "cost" or the inefficiency caused by the tax system can be estimated. However, in order to make recommendations about optimal tax design, the adviser must be aware of the preferences of the government (or the society) involved. Without this information, the advice rendered represents only the adviser's preferences, which need not be the same as those of the government seeking advice. Hence, there is interest in examining the possibility of overcoming this difficulty by statistical analysis.

The aim of this paper is to suggest a method that enables the user to identify commodities that a large class of social-welfare functions would reveal as being worth subsidizing or taxing. If such commodities can be identified, then the task of advising the government on optimal directions of tax reform will be rendered nearly value-free. Alternatively, nearly all social-welfare functions would indicate that the taxation of one commodity should be reduced in favor of heavier taxation on another.

The specific question this paper addresses is the following. Assume that the social evaluation of the marginal utility of income

*Hebrew University, Jerusalem 91905, Israel; University of Michigan, Ann Arbor, MI 48109. The first version of the paper was written while Yitzhaki served as a consultant at the World Bank. The World Bank does not accept responsibility for the views expressed herein, which are our own and should not be attributed to the World Bank or to its affiliated organizations. The findings, interpretations, and conclusions are the result of research supported by the World Bank; they do not necessarily represent its official policy. We thank Reza Firuzabadi for research assistance and Joram Mayshar, Wayne Thirsk, and three anonymous referees for their comments.

is positive and declining. The government plans to make an equal-yield change in its commodity-tax system by subsidizing one commodity and taxing another commodity by one additional dollar. Is it possible to identify two such commodities such that welfare increases for all additive concave social-welfare functions? If such situations can be identified, a preferred direction of tax reform can be located without detailed knowledge of preferences regarding interpersonal transfers. If such commodities *cannot* be found, it will clearly be impossible to make recommendations in the absence of further information about the governing social-welfare function.

The methodology that enables us to answer such questions was originally developed in the finance literature, where it is referred to as the "second-degree stochastic dominance" criterion. The main idea is to rank portfolios according to their expected utility, such that the investigator assumes only that the marginal utility of income is nonnegative and nonincreasing. Based on these assumptions, rules for ranking prospects have been developed (e.g., Josef Hadar and William Russell, 1969; Giora Hanoach and Haim Levy, 1969).

Our intention is to use the methodology of stochastic dominance to evaluate changes in the taxation of commodities. As has been demonstrated by Anthony Atkinson (1970), there is a formal similarity between the ranking of income distributions and the ranking of prospects. Hence, the use of stochastic-dominance rules in welfare economics is a natural development following from Atkinson's observation. However, since taxation (in particular, commodity taxation) affects social welfare in a slightly different way than the effect of a portfolio on the utility function, several changes must be made in the methodology. We refer to these adapted rules as welfare dominance.¹ The major changes are the following:

- (a) In the finance literature, the main interest is to rank portfolios, the analogy to

which, in our study, is the ranking of income distributions. Our goal, though, is the ranking of (taxes on) commodities, expenditure on which is only one component of total income. Therefore, we have to use conditional stochastic-dominance rules. The analogy in the finance literature is the issue of whether asset A dominates asset B, given that the investor also has to hold portfolio C. In the case of welfare dominance, the same formal question can be interpreted as whether subsidizing expenditure on commodity A financed by a tax on commodity B improves social welfare, given that the income distribution is C.

- (b) We will be interested in dominance at the margin. The analogy to finance is whether a small increase in the share of asset A at the expense of asset B (given that the rest of the portfolio held is C) increases expected utility. In the case of taxation, the same formal question is now whether a small decrease in the tax on commodity A financed by a small increase in a tax on B (given that the income distribution is C) increases social welfare.
- (c) In welfare economics, one has to take into account the efficiency implications of the reform. This means that an increase of one dollar in the revenue from the tax on one commodity enables the government to subsidize another commodity by an amount that may be larger or smaller than one dollar, depending on efficiency considerations.

As we show in the next section, this question can be answered by comparing shifted concentration curves. The concentration curve is a diagram similar to the Lorenz curve. On the horizontal axis the households are ordered according to their income, while the vertical axis describes the cumulative percentage of the total expenditure on a specific commodity that is spent by the families whose incomes are less than or equal to the specified income level. The concentration curve, like the Lorenz curve, passes through the origin; but, unlike the

¹This term was coined by Anthony Shorrocks (1983).

Lorenz curve, it need not always be increasing, and its curvature depends on the structure of the income elasticity of the commodity. In particular, if the curve is convex (concave) to the origin, then the income elasticity is negative (positive).² Efficiency implications of the reform result in shifts of the concentration curves.

We demonstrate that if the (shifted) concentration curve of one commodity is above the (shifted) concentration curve of another commodity, then the first commodity dominates the second in the sense that a small tax decrease in the first accompanied by a tax increase in the second (with revenue remaining unchanged) increases social welfare. However, if the (shifted) concentration curves intersect, then it is impossible to show dominance. In other words, if and only if concentration curves do not intersect will all additive social-welfare functions show that the tax change increases welfare.³ We refer to these rules as "marginal conditional stochastic dominance" rules (hereafter MCSD rules). In the remainder of the paper, we will use the term "dominate" when one commodity has marginal conditional stochastic dominance over another.

The next section provides an intuitive proof for MCSD rules. In the second section, additional insight is gained by relating these rules to a methodology based on the decomposition of the Gini coefficient. Section III uses data from Israel to illustrate the methodology. The paper concludes with suggestions for further research.

I. Intuitive Derivation of the Methodology⁴

Assume that tax policies are evaluated according to an additively separable social-welfare function.⁵ All that is known about the social-welfare function is that the social evaluation of the marginal utility of income is positive and declining. Formally,

$$(1) \quad W = \sum_h w[v^h(y_h, P_1, \dots, P_n)]$$

where y is income and the P 's are prices, v is the indirect utility, and w is the social evaluation of the utility of individual h . Individuals may have different utility functions. However, we assume that if $y_i = y_j$ then $w[v^i(y, P_1, \dots, P_n)] = w[v^j(y, P_1, \dots, P_n)]$ and that the social evaluation of the marginal utility of income, denoted β , is a function of y (and prices) only.⁶ That is, $\beta(y) = (\partial w / \partial v^h)(\partial v^h / \partial y)$ is a positive and declining function of y with $\partial \beta / \partial y < 0$. We refer to β as marginal welfare.

Now suppose the government is considering a small increase in the tax on commodity t and a small decrease in the tax on commodity s , so that total revenue does not change. Let x_i^h denote the consumption of commodity i by the h th individual, where individuals are arranged in nondecreasing order of income, and let X_i denote total consumption of commodity i . Since the change in revenue raised is zero, there is a link between the change in the prices of

²For a detailed analysis of the curvature of concentration curves, see Nanak Kakwani (1980) and Yitzhaki and Ingram Olkin (1988).

³It is worth noting that concentration curves have long been used to describe the progressivity of taxes (see, e.g., Daniel Suits, 1977a,b; Charles Clotfelter, 1979; John Formby and David Sykes, 1984; Formby et al., 1986a,b; Kakwani, 1977a,b, 1980; Thomas Calmus, 1981; Pak-Wai Liu, 1985; Donald Kiefer, 1984). However, the use of concentration curves to identify welfare-improving tax changes is new. As will be shown later, this interpretation requires several adjustments to the way concentration curves are defined.

⁴In Yitzhaki and Olkin (1988), MCSD rules are developed in the context of portfolio analysis. Since the proofs are similar to those required in this paper, we refer the interested reader to that paper. In this section, an intuitive proof is given.

⁵The results of the paper also apply to an increasing, S-concave, social-welfare function. The requirement for an additive social-welfare function enables us to present intuitive proofs.

⁶Income is defined here in a broad sense, so that it can be interpreted as income per capita, income divided by the poverty ratio, etc. All that is required is that, if the income of one individual is higher than the income of another individual, then the welfare of the first individual is higher than the welfare of the second.

commodities t and s . Let R be the tax revenue. Assuming that there are K taxed (subsidized) commodities and that producer prices are normalized to 1, then:

$$(2) \quad R = \sum_k \tau_k X_k$$

where τ denotes the tax (or subsidy) rate. Reducing the tax on (subsidizing) commodity s and increasing the tax (reducing subsidy) on commodity t while keeping tax revenue unchanged implies that

$$(3) \quad dR = 0 = dP_t X_t \alpha_t + dP_s X_s \alpha_s$$

where $\alpha_i = 1 + \sum_k (\tau_k / X_i) (\partial X_k / \partial \tau_i)$. P_i refers to the consumer price i , so that $P_i = 1 + \tau_i$. The term α_i reflects the revenue effect of a change in τ_i ; in general, it depends on all tax rates and the properties of the demand functions. David Wildasin (1984) and Joram Mayshar (1988) interpret α_i as the marginal social cost of raising one dollar of revenue by taxing the i th commodity. Rearranging terms, equation (3) can be written as

$$(4) \quad dP_t = - (X_s / X_t) \alpha_{st} dP_s$$

where $\alpha_{st} = (\alpha_s / \alpha_t)$ can be interpreted as the ratio of the marginal costs of public funds. If $\alpha_{st} < 0$, then both prices can be decreased while holding revenue constant.⁷ Since this is a trivial case, it is assumed that $\alpha_{st} > 0$. We consider other important values of α_{st} in the next section.

Now consider the change in the utility of individual h when tax rates are changed. Denote his consumption of commodities s and t by x_s^h and x_t^h . Omitting the index h (for simplicity of presentation), the change

in his utility is

$$(5) \quad dv = v_s dP_s + v_t P_t \\ = -v_y (x_s dP_s + x_t dP_t)$$

where v_s is the derivative of $v(\cdot)$ with respect to P_s . The second equality is a result of Roy's identity. Substituting equation (4) into (5) we get

$$(6) \quad dv = -v_y [(x_s / X_s) - \alpha_{st} (x_t / X_t)] X_s dP_s.$$

Since v_y and X_s are positive and dP_s is negative, it is clear that individual h gains from the reform if the expression in brackets in (6) is positive. If this expression is nonnegative for all h , then the tax reform is a Pareto improving tax reform.⁸

A necessary condition, namely, that all additive concave social-welfare functions should show that the suggested reform is welfare-increasing, is that the reform does not worsen the utility of the poorest individual. (Otherwise a Rawlsian decision-maker will judge that the suggested reform decreases welfare.) The effect of the reform on the welfare of the poorest individual in the society can be seen by applying equation (6) to evaluate the change for this individual. The welfare of the poorest individual does not worsen if

$$(7) \quad (x_s^1 / X_s) - \alpha_{st} (x_t^1 / X_t)$$

is nonnegative. This expression is the difference between the poorest's individual's share of consumption of commodity s and

⁷To assume $\alpha_{st} < 0$ means that we are on the declining side of the Laffer curve. Alternatively, to require that α_s and α_t are both positive is to assume that an increase in the tax on each commodity increases tax revenue. Tatsuo Hatta and John Haltiwanger (1986) define such commodities as revenue-increasing commodities.

⁸The conditions for Pareto improving tax reforms are presented in Avinash Dixit (1975), Dixit and K. J. Munk (1977), Roger Guesnerie (1977), W. Erwin Diewert (1978), and Serge Wibaut (1987). The concept of a Pareto improving tax reform is more restrictive than the concept of first-degree stochastic dominance, because every individual must gain from the reform. Therefore, it must be based on efficiency gains. In contrast, our method is based on rules of second-degree stochastic dominance, which allow some individuals to lose due to the reform, provided that social welfare increases.

his share of commodity t multiplied by α_{st} . Hence, a necessary condition for all social-welfare functions to show an increase in social welfare is that the share of the expenditure of the poorest individual on commodity s should be higher than his share in the expenditure on commodity t multiplied by a constant.

Having established the condition for a socially acceptable reform with regard to the poorest individual, we next check the necessary condition applying to the poorest and next-to-poorest individuals together. By assumption, the marginal utility of income of the second individual is lower than the marginal utility of income of the first individual. Thus, another necessary condition is that the gain in real income for the first individual is higher than the loss, if any, for the second individual. By repeating the same considerations that led to (7), we obtain

$$(8) \quad [(x_s^1 + x_s^2)/X_s] - \alpha_{st}[(x_t^1 + x_t^2)/X_t] > 0.$$

The set of necessary conditions that applies to each group of individuals from 1 to N is thus

$$(9) \quad \left\{ \left[\sum_{h=1}^k x_s^h / X_s \right] - \alpha_{st} \left[\sum_{h=1}^k x_t^h / X_t \right] \right\} \geq 0$$

for $k = 1, \dots, N$.

Condition (9) has a straightforward interpretation. The expression in braces is the difference between the height of the relative concentration curve of commodity s and the height of the relative concentration curve of commodity t multiplied by a constant. The condition implies that for all additive social-welfare functions to show that this tax change increases social welfare it is necessary that the concentration curve of commodity s with respect to income is at least as high as the concentration curve of commodity t multiplied by a constant at each point of the income distribution. In Appendix A, it is shown that condition (9) is also sufficient for welfare dominance. We summarize the results in the following theorem.

THEOREM 1: *A necessary and sufficient condition for commodity s to dominate commodity t is that the concentration curve of s is at least as high as the concentration curve of t multiplied by α_{st} at each point of the income distribution.*

To explain this finding, it is instructive to construct what we refer to as the "difference in concentration curves" curve [hereafter the DCC(F) curve, where F is the cumulative distribution of income].⁹ The horizontal axis of this curve represents the cumulative distribution of income, while the vertical axis represents the difference between the concentration curve of commodity s and the concentration curve of commodity t multiplied by a constant, α_{st} . In other words, on the vertical axis we present the share of total expenditure on s minus α_{st} times the share of total expenditure on t for the poorest F families. Although all concentration curves start at (0,0) and end up at (1,1), all the DCC curves start at (0,0) and end up at $(0, 1 - \alpha_{st})$. The theorem states that, if the DCC(F) curve is everywhere above the horizontal axis, then s dominates t ; if it is entirely below the horizontal axis, then t dominates s . If the curve crosses the horizontal axis, neither s nor t dominates the other.

The DCC curve enables us to present in one figure both the efficiency gain from the reform and the way the gain is distributed among income groups. As shown in Yitzhaki and Wayne Thirsk (1990), the DCC curve describes the cumulative gain in real income for the poorest F households in society. The DCC curve is increasing (decreasing) if the individual at F gains (loses) from the reform.¹⁰ Since the curve reflects both ef-

⁹For simplicity of exposition, it is assumed that distributions are continuous. F denotes the cumulative distribution of the population, and its range is $[0,1]$.

¹⁰Note, however, that even if the DCC curve is increasing everywhere, it does not form a sufficient condition for a Pareto improving tax reform. If several individuals have the same income, the curve increases if the group (and not necessarily every individual in it) gains.

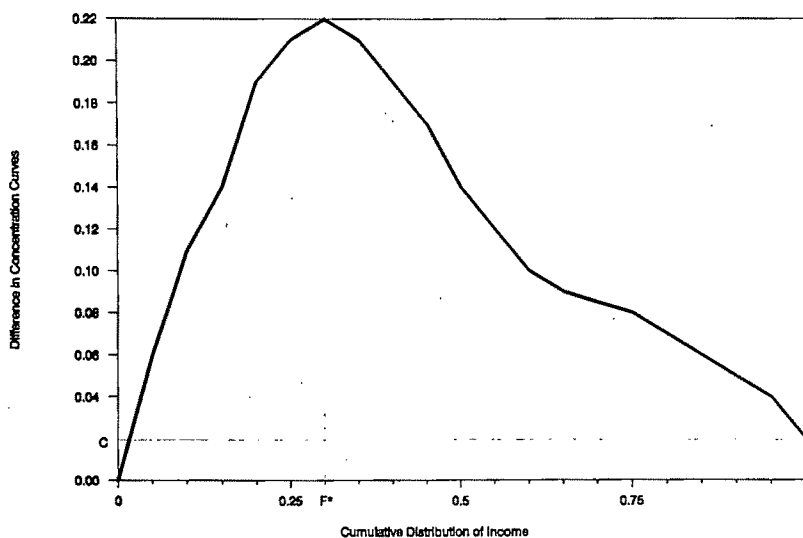


FIGURE 1. THE DCC CURVE

efficiency and distributive gains, it is convenient to separate the effects of these two causes.

The efficiency gains of a tax reform can be seen by concentrating on $DCC(1)$. At this point, the DCC curve describes whether the cumulative gain to the society is positive, zero, or negative. It is easy to see that $DCC(1)$ is larger (smaller) than 0 if and only if α_{st} is smaller (larger) than 1. In a one-consumer economy, the question of whether society gains from the reform collapses into whether $DCC(1)$ is greater than 0, or alternatively, whether α_{st} is less than 1. It can be shown that the condition $\alpha_{st} < 1$ is identical to the condition for a welfare-improving tax reform presented by Hatta and Haltiwanger (1986) in a one-consumer economy.¹¹ Therefore, we can interpret the condition $\alpha_{st} < 1$ as indicating whether the excess burden declines as a result of the reform. However, in a multiperson economy, a reform can be welfare-improving even if $\alpha_{st} = 1$,

the case in which there is no efficiency gain or loss. This can occur because the reform improves the distribution of welfare. In this case, the DCC curve will be above the horizontal axis, with $DCC(1) = 0$. If $DCC(1) < 0$ (i.e., $\alpha_{st} > 1$), then condition (9) cannot hold for all h , because it cannot hold for $h = N$. Thus, dominance is impossible. The intuitive explanation is that when $\alpha_{st} > 1$ the tax-policy change has increased the excess burden of the tax system. Therefore a decision-maker using a social-welfare function with (almost) constant marginal utility of income will oppose such a tax reform, which reduces the sum of all incomes. Thus, when $\alpha_{st} > 1$, agreement among all possible policymakers is impossible without further restricting the social-welfare function. If $\alpha_{st} < 1$, then a Pareto improving tax reform is possible. This would occur if the DCC curve increases everywhere, and every individual gains from the reform.

Distributive gains can be seen from the curvature of the DCC curve. If the DCC curve is increasing (decreasing), then the individual at that point of the income distribution is gaining (losing) from the reform, and the greater the slope, the larger the gain.

Figure 1 is an example of a DCC curve. The curve increases from 0 to F^* , which

¹¹It can be shown that the condition $\alpha_{st} < 1$ is identical to $N^{ab} > 0$ in lemma 1 of Hatta and Haltiwanger (1986 p. 307). They investigate the conditions under which $\alpha_{st} < 1$. If lump-sum taxes are available, then it is easy to see that $\alpha_{st} < 1$ is a sufficient condition for a Pareto improving tax reform. Appropriate lump-sum taxes ensure that the curve rises everywhere.

means that the lowest F^* individuals benefit from the reform, and it declines afterward, meaning that the top $1 - F^*$ individuals lose from the reform. The overall gain is equal to C .

Note that condition (9) does not constitute a complete ordering of revenue-neutral (commodity) tax changes. That is, if two concentration curves intersect, it will be impossible to find a small change in the taxation of the goods s and t such that all social-welfare functions can be judged to increase welfare. A complete ordering would require investigating the concentration curves further, to see whether restricting the class of social-welfare functions enables us to classify policies completely. This issue is discussed in the next section.

Although the ordering is not complete, it is clearly transitive: if commodity A dominates commodity B and commodity B dominates C, then A dominates C.¹² This property enables us to establish an ordering within the subgroup of commodities in which dominance is found.

Until now, we have been interested in an increase in the subsidy to one commodity, financed by an increase in the tax on another commodity. The same methodology can be used to ask whether an increase in a subsidy that is financed by a proportional income tax increases welfare. In this case, we treat total income as a commodity and look at the difference between the concentration curve for commodity s and the Lorenz curve. If the $DCC_{s,0}$ (where the index 0 indicates the Lorenz curve), is always positive, then commodity s dominates the proportional tax.

Finally, two important points are worth mentioning. As discussed above, concentration curves have often been used in the literature to describe the progressivity of taxes. However, the curves used in this pa-

per differ slightly from the standard use of concentration curves. In this paper, the curves are defined for after-tax income and are used to describe the progressivity of small changes in taxes. The reason for this modification is the following. In order to be consistent with welfare dominance, the ordering on the horizontal axis should be the ordering of the social evaluation of the utility of individuals. Since individuals with the same income may consume different quantities of commodities, changes in taxes may affect the ordering. The possibility of reordering which results in horizontal inequities restricts the use of concentration curves to cases involving small changes in taxes in order to indicate the preferred direction of the tax reform. In the case of a pure income tax, there is no problem of reordering, and therefore one can use Lorenz curves to order income taxes (see Roger Latham, 1988).

II. A Class of Easily Computable Necessary Conditions for Welfare Dominance

As noted above, the MCSD rules do not form a complete ordering across the taxable commodities. Moreover, we suspect that many commodities cannot be ordered by this method. In these cases, one may wish to investigate further restrictions on the set of possible welfare functions, such as rules of third-degree stochastic dominance.¹³ Applying such rules does not, however, ensure complete ordering in all cases. Therefore, their use may increase the set of commodities over which an ordering is defined, but it does not eliminate the problem of incomplete ordering.

¹²To see this, note that if the concentration curve of A is everywhere above the concentration curve of B, and the concentration curve of B is everywhere above the concentration curve of C, then it is clear that the concentration curve of A is everywhere above the concentration curve of C.

¹³For a definition of third-degree stochastic dominance, see G. A. Whitmore (1970). For an application in welfare economics, see Shorrocks and J. E. Foster (1987). It can be shown that a necessary and sufficient condition for third-degree marginal conditional stochastic dominance is that the area between the DCC curve and horizontal axis (i.e., the integral of DCC) is positive everywhere and that $DCC(1) > 0$. However, conditions of third-degree stochastic dominance are beyond the scope of this paper.

A procedure that does ensure complete ordering is to restrict ourselves to necessary conditions for welfare dominance. This is the case in which the analysis of tax reform is carried out using the Gini coefficient (Yitzhaki, 1987). This procedure provides a complete ordering that will never be rejected by any additive concave social-welfare function. It also is useful for making MCSD rules operational, because investigating the possibility of welfare dominance requires that $n(n-1)/2$ DCC curves be analyzed. As we show later, the use of necessary conditions decreases the number of curves that have to be investigated, and the procedure can be performed with a simple regression program.

The area below the 45° line minus the area below the concentration curve is defined as one-half of the concentration ratio. As shown in Yitzhaki and Olkin (1988), this area is also equal to

$$(10) \quad C_i = \text{Cov}[X_i, F(Y)] / m_i$$

where C_i is one-half of the concentration ratio, m_i is the mean expenditure on commodity i , and $F(Y)$ is the cumulative distribution of income. In other words, the concentration ratio is equal to twice the covariance between the expenditure on commodity i and the cumulative distribution of income divided by the mean expenditure on commodity i . Hence, the area between the concentration curve of commodity s and the concentration curve of commodity t (i.e., between the DCC_{st} curve and the horizontal axis) is¹⁴

$$(11) \quad C_s - \alpha_{st} C_t = \text{Cov}[X_s, F(Y)] / m_s \\ - \alpha_{st} \text{Cov}[X_t, F(Y)] / m_t.$$

By dividing and multiplying equation (11) by

$\text{Cov}[Y, F(Y)]$ and m_Y we can rewrite (11) as

$$(12) \quad \int_0^1 \text{DCC}_{st}(F) dF \\ = \{(b_s/S_s) - \alpha_{st}(b_t/S_t)\} G_Y$$

where G_Y is the Gini coefficient of income, S_i is the share of the expenditure on commodity i , and

$$(13) \quad b_i = \text{Cov}[X_i, F(Y)] / \text{Cov}[Y, F(Y)]$$

is a nonparametric estimator of the slope of the regression line of X_i on Y (Yitzhaki, 1987). In our context, b_i is a weighted mean of the marginal propensity to spend on commodity i . As argued in Yitzhaki (1987), b_i/S_i can be interpreted as the weighted average income elasticity of commodity i . Hence equation (12) tells us that the sign of the area below the DCC_{st} curve is determined by the difference between the weighted average income elasticities of the commodities. For commodity s to dominate commodity t , DCC_{st} must be positive; thus, it is clear that a necessary (but, of course, not sufficient) condition for welfare dominance is that the income elasticity of commodity s be lower than that of commodity t .

The use of the Gini coefficient means that we use a specific welfare function with a specific weighting scheme. However, one can use other weighting schemes. A family of weighting schemes that is based on the same argument is offered by the extended Gini (Yitzhaki, 1983). The extended Gini is a weighted integral of the area between the 45° line and the Lorenz curve. The formula for the extended Gini is

$$(14) \quad G(\nu) = -\nu \text{Cov}\{Y, [1 - F(Y)]^{\nu-1}\} / m_Y \\ \nu > 1$$

where ν is a parameter chosen by the investigator. The extended Gini is similar to the Gini coefficient but uses a different weighting scheme. The Gini is a special case of the extended Gini, where ν is 2. The higher is

¹⁴The derivation of this equation and that of equation (15) appear in Appendix B.

ν , the greater is the emphasis on the bottom of the income distribution.¹⁵

The above analysis using the Gini can be carried out using the extended Gini. In Appendix B, we show that a decrease in the consumer price of commodity s , financed by an increase in the consumer price of commodity t , decreases the extended Gini inequality index, if

(15)

$$\int_0^1 [\Theta_s(F) - \alpha_{st}\Theta_t(F)](1-F)^{\nu-2} dF > 0$$

where $\Theta(F)$ is the concentration curve. Since $\Theta_s - \alpha_{st}\Theta_t$ is the DCC_{st} curve, the analysis of tax reform with the extended Gini coefficient provides additional necessary conditions for welfare dominance. If commodity s dominates commodity t , then a shift from taxing commodity s to commodity t must decrease the extended Gini inequality index for all ν , including the standard Gini case where $\nu = 2$. These necessary conditions are useful in the empirical investigation of welfare dominance because they are fairly easy to calculate (see Robert Lerman and Yitzhaki, 1989) and can be used to identify those pairs of commodities for which welfare dominance is possible.

III. An Illustration with Israeli Data

This section illustrates the methodology of welfare dominance using cross-sectional data on the consumption of subsidized commodities in Israel. The data set is the Survey of Family Expenditure (1979/1980) conducted by the Israel Central Bureau of Statistics. This survey consists of a random stratified sample of 2,271 urban households. Since we are interested in the level of economic well-being, the concept of income per standard adult is used.¹⁶ The households

are ordered according to net total income per standard adult, where net total income is defined as monetary income plus imputed income from ownership of housing and vehicles minus income and social security taxes.¹⁷

Before presenting the results, two technical points are worth making. In general, if there are n commodities, one would have to plot $n(n-1)/2$ pairs of concentration curves to investigate the existence of welfare dominance. Since cross-sectional samples usually contain thousands of observations, this can clearly become a cumbersome procedure when more than a few commodities are being studied. As suggested above, comparisons of the magnitude of the weighted income elasticities, according to the Gini or extended Gini coefficient, yield necessary conditions for dominance, thus reducing the number of comparisons of curves needed. The second issue arises because of the use of a sample instead of the whole population. As shown by Charles Goldie (1977), sample-based Lorenz curves do converge to the population Lorenz curve, but it is clear that our results may be affected by the sample variability. One way to reduce the sample variability is to average observations. However, some information is lost by averaging. Hence, in what follows we plot concentration curves based on the whole sample and also report the results when concentration curves are based on averaging consecutive pairs of observations. We assume that $\alpha_{st} = 1$, implying that reform neither increases nor decreases the excess burden.

size of the household on consumption needs. The scale that is used is the following: single = 1.25 standard adults; a couple without children = 2.0; a couple with one child = 2.65; with two children = 3.2; with three children = 3.75; with four children = 4.2; and 0.4 for each additional child. This scale is used in many official publications of the National Insurance Institute and the Central Bureau of Statistics in Israel.

¹⁵See Yitzhaki (1983) for a discussion of the properties of the extended Gini. See Wilhelm Pfähler (1987) for a discussion of a general class of progressivity measures.

¹⁶The concept of a standard adult is an equivalence scale intended to take into account the effect of the

¹⁷The sample is a stratified one, which means that each household in the sample represents a different number of households in the population. The equations given in this paper have therefore been adjusted in a straightforward manner to account for this (see Lerman and Yitzhaki, 1989).

TABLE 1—INCOME ELASTICITIES OF SEVERAL SUBSIDIZED COMMODITIES
IN ISRAEL DURING 1979/1980

Gini parameter	$\nu = 1.5$	$\nu = 2$	$\nu = 4$
Cooking oil	-0.13	-0.14	-0.11
Bread	-0.06	-0.07	-0.09
Sugar	0.07	0.06	0.08
Public transportation	0.12	0.15	0.25
Water for household consumption	0.31	0.30	0.33

Source: Authors' calculations from The Survey of Family Expenditure 1979/1980, Central Bureau of Statistics, Israel.

Table 1 presents weighted average income elasticities, estimated by using several variants of the extended Gini coefficient. As can be seen, there are two inferior commodities (bread and cooking oil), while the other commodities are normal.

As mentioned above, the higher is ν the greater is the weight given to the lower portion of the income distribution. Hence, we can conclude from the table that the income elasticity of public transportation declines as income increases, while bread tends to be less inferior the higher the income level. Since the income elasticity of cooking oil is lower than the income elasticity of bread for all the values of ν we studied, the necessary conditions for cooking oil to dominate bread are met, and it is reasonable to check whether cooking oil dominates bread. By the same reasoning, we conclude from Table 1 that, if there is dominance, the ordering must be: cooking oil, bread, sugar, public transportation, and finally water for household consumption. It is impossible that a good will be dominated by another good lower in the table.

Figure 2 presents the curve of the difference between the concentration curves (DCC curve) of cooking oil and bread. As expected from Table 1, the area above 0 on the x-axis is larger than the area below it. (The difference between the areas is equal to the income elasticity of cooking oil minus that of bread calculated by the Gini index.) Hence, we can see that cooking oil is more inferior than bread when the weighting

scheme of the Gini is used. However, the curve intersects 0 on the x-axis, implying that for the lower two deciles of the income distribution bread has a lower income elasticity than cooking oil, while for the other deciles bread has a higher income elasticity. If the social-welfare function is concave only for the lower two deciles, then taxation policy should shift to subsidize bread at the expense of cooking oil. The conclusion is that cooking oil does not dominate bread.

Figure 3 presents the DCC curve between bread and water for household consumption. As can be seen, the curve is always above 0 on the x-axis, and hence bread dominates water.¹⁸ A small increase in the tax on water that finances a small increase in the subsidy to bread will be welfare-improving for all concave social-welfare functions.

¹⁸ A close examination of the individual observations reveals that the curve intersects the x-axis for the very last observation, so that the share of the richest household in the sample in total expenditure on bread is higher than its share in the overall expenditure on water. If this sample exactly portrayed the population, a social-welfare function that is linear over the whole range of the distribution and concave between the second-richest family and the richest one would show that subsidizing bread at the expense of water would be welfare-decreasing. However, we suspect that this is a result of a sampling error. If we take the average of two consecutive observations, this proviso disappears. Note, however, that the question of sampling variability is not addressed.

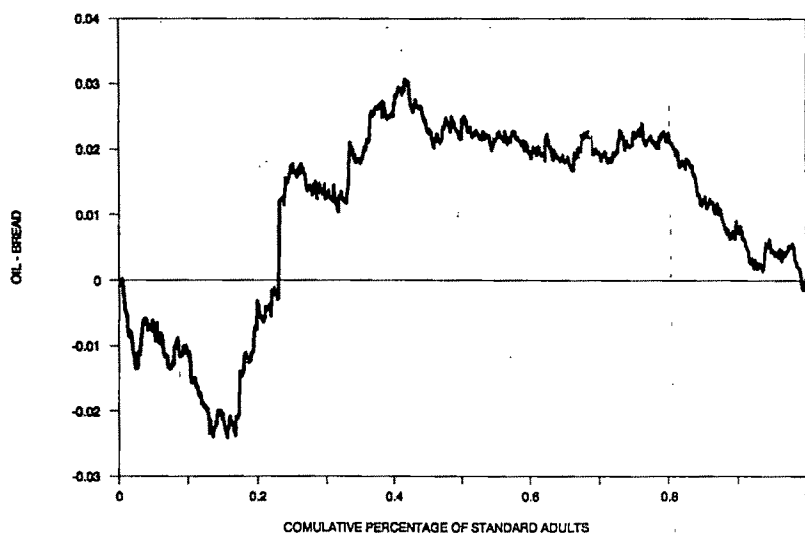


FIGURE 2. THE DIFFERENCE BETWEEN CONCENTRATION CURVES OF OIL AND BREAD

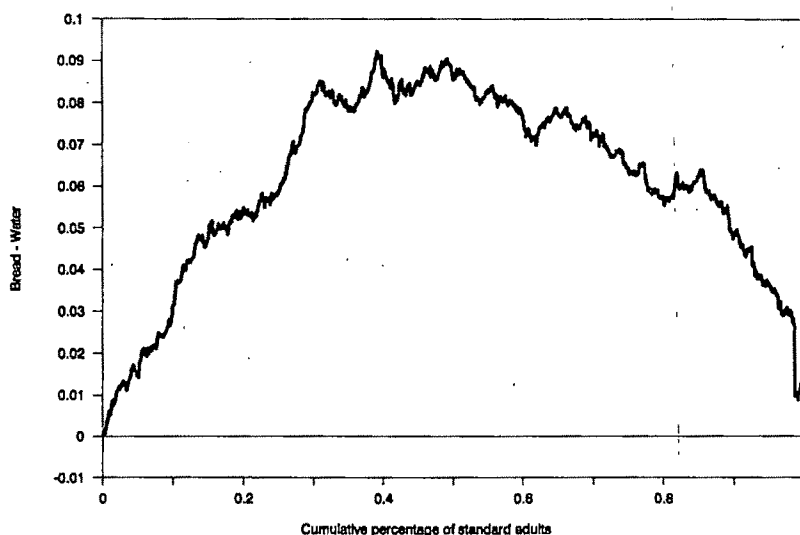


FIGURE 3. THE DIFFERENCE BETWEEN CONCENTRATION CURVES OF BREAD AND WATER

Figure 4 presents the DCC of bread and public transportation. Except for one observation, bread dominates public transportation. If we plot the DCC averaging every two consecutive observations, then bread dominates public transportation with no exceptions. This phenomenon illustrates just

how strong the condition of MCSD is. Even if the entire population is studied, the consumption patterns of the poorest and the richest members of society are critical in determining the existence of dominance.

Figure 5 presents the DCC of bread minus the Lorenz curve. This figure is in-

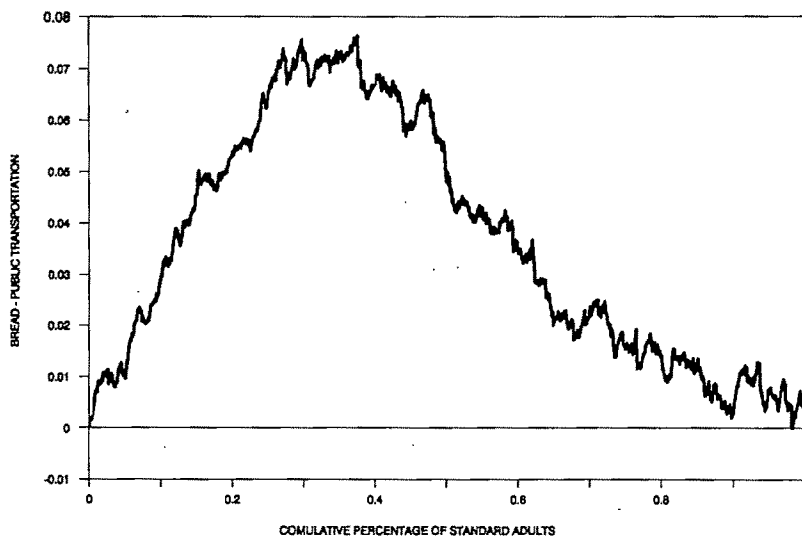


FIGURE 4. THE DIFFERENCE BETWEEN CONCENTRATION CURVES OF BREAD AND PUBLIC TRANSPORTATION

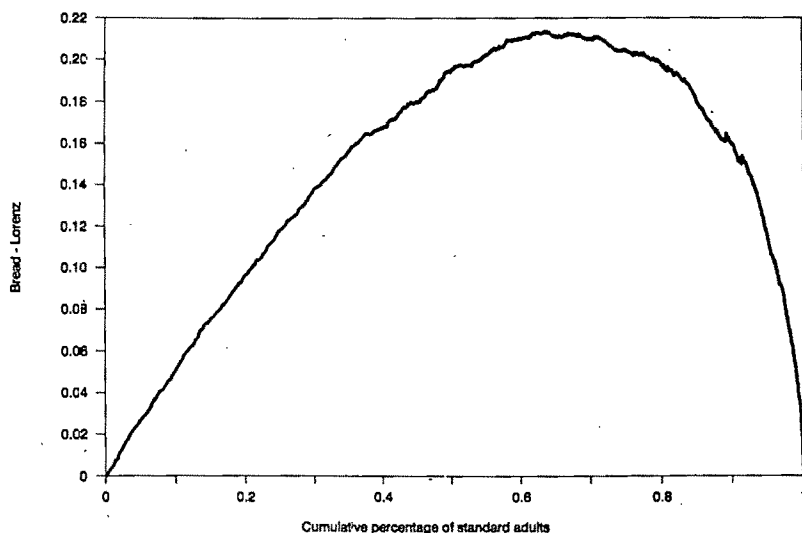


FIGURE 5. THE DIFFERENCE BETWEEN THE CONCENTRATION CURVE OF BREAD AND THE LORENZ CURVE

tended to show whether an increase in the subsidy on bread, financed by an increase in a proportional income tax, increases welfare for all additive concave Paretian social-welfare functions. Because the DCC curve is above 0 on the x-axis over the entire income range, the answer to this question is yes.

IV. Conclusions

Because the exact properties of social-welfare functions are not known and are likely to remain so, it would be valuable to make judgments about potential tax reforms that depend only on relatively noncontroversial characteristics of the social-welfare

function. The methodology provided in this paper is a step in this direction. It derives the conditions (called marginal conditional stochastic dominance rules) that must hold for all individuals with additive Paretian concave social-welfare functions to agree that an increase in the subsidy on one commodity, financed by an increase in the tax on another commodity (or a proportional income tax), increases social welfare. If this reform does not increase the excess burden, then all individuals will agree on the preferred direction of tax reform. An inspection of Israeli data suggests that these conditions are quite commonly observed in practice.

One direction for future research is to apply this methodology to a weighted combination of commodities. In this case, a set of changes in taxes and subsidies can be compared to another set in order to see whether all individuals can agree on the welfare implications of a particular tax reform. In the present setting of MCSD rules, the permissible set of welfare functions includes welfare functions that are almost linear (constant marginal utility). The permissibility of such welfare functions means that it will be impossible to find unanimous preference for any redistribution that increases efficiency costs. Therefore, only by limiting the set of the admissible welfare functions will it be possible to find MCSD rules for costly redistribution. One possible way of doing this is to restrict the set such that only welfare functions with some minimum concavity are included.

APPENDIX A

In this appendix, we prove that if the concentration curve of s is at least as high as the concentration curve of t times α_{st} , then it is impossible to find a welfare function that shows that subsidizing t (and taxing s) increases welfare.

Formally, if

$$(9) \quad \left\{ \left[\sum_{h=1}^k x_s^h / X_s \right] - \alpha_{st} \left[\sum_{h=1}^k x_t^h / X_t \right] \right\} \geq 0$$

for all k

or, in its continuous version,

(A1)

$$\int_0^y [(x_s(z)/X_s) - \alpha_{st}(x_t(z)/X_t)] f(z) dz \geq 0$$

for all y

where $f(z)$ is the density function, then it is impossible to find a welfare function that shows that decreasing the price of t and increasing the price of s while keeping revenue constant increases welfare.

To prove this assertion, we first assume that such a welfare function exists. Then,

$$(A2) \quad \int w_v(v_s dP_s + v_t dP_t) f(y) dy > 0$$

where $w_v = \partial w / \partial v > 0$. Using Roy's identity and (4), we find that (A2) is equal to

$$(A3) \quad \int -\beta(y) [x_s(y) dP_s + x_t(y) dP_t] f(y) dy$$

$$= -dP_s \int \beta(y) [(x_s(y)/X_s) - \alpha_{st}(x_t(y)/X_t)] f(y) dy \geq 0$$

where $\beta(y) = w_v v_y$ is the marginal welfare. Since dP_s is positive, the integral is nonpositive. Since $\beta(y)$ is nonnegative and nonincreasing with y , we can use the following lemma from Yitzhaki (1982 p. 179), which completes the proof.

LEMMA: Let $V_y(y)$ be a nonnegative and nonincreasing function and let $B(y)$ be a function with the following property: $\int_0^y B(z) dz \geq 0$ for all y . Then $\int_0^y V_y(z) B(z) dz \geq 0$ for all y .

Now, if one lets $B(z) = [(x_s(z)/X_s) - \alpha_{st}(x_t(z)/X_t)] f(z)$, then by (A1) the requirements for the lemma are fulfilled. Hence, (A3) is contradicted.

APPENDIX B

In this appendix we prove the following results:

- (a) the derivative of the extended Gini of overall income with respect to equal-revenue tax changes is equal to the difference in the (weighted) concentration ratio of the commodities involved in the tax changes;
- (b) the (weighted) concentration ratio is equal to the (weighted) area between the 45° line and the concentration curve.

Properties (a) and (b) together with equation (4) ensure that a necessary condition for one commodity to dominate another is that the extended Gini of inequality declines as a result of the changes in taxes.

PROOF OF PROPERTY (a):

For simplicity of presentation, we assume that the distributions are continuous and that only two commodities are involved. Let

$$(B1) \quad y = P_1 x_1 + P_2 x_2$$

where $P_i = 1$ is the price of commodity i , y denotes income, and x represents commodities. The extended Gini inequality index is

$$(B2)$$

$$G_y(\nu) = -\nu \text{Cov}(y, [1 - F(y)]^{\nu-1}) / m_y$$

where m denotes the mean.

The derivative of $G_y(\nu)$ with respect to P_1 , evaluated at $P_1 = 1$, can be interpreted as the effect of a change in the tax on commodity 1 on the extended Gini. Note that P_1 affects both the numerator and the denominator of (B2). Let us start with the derivative of the numerator.

Let $A(y) = 1 - F(y)$. Note that

$$E_y[A^{\nu-1}(y)] = \int_0^\infty [1 - F(y)]^{\nu-1} f(y) dy$$

and by transformation of variables with $F =$

$F(y)$ we get

$$(B3) \quad E_y[A^{\nu-1}(y)] = \int_0^1 (1 - F)^{\nu-1} dF = 1/\nu.$$

Using (B3), the numerator of (B2) can be written as

$$\text{Cov}[y, A^{\nu-1}(y)] = \int_0^\infty y[\nu A^{\nu-1}(y) - 1] f(y) dy$$

and again using the same transformation of variables where $F = F(y)$

$$(B4) \quad \text{Cov}[y, A^{\nu-1}(y)] = \int_0^1 y(F) [\nu(1 - F)^{\nu-1} - 1] dF$$

where $y(F) = \inf\{y | F(y) > F\}$ is the inverse of the cumulative distribution. Differentiating (B4) while using (B1) yields

$$(B5) \quad \partial \text{Cov}[y, A^{\nu-1}(y)] / \partial p_1 = \int_0^1 x_1(F) [\nu(1 - F)^{\nu-1} - 1] dF$$

and, by reversing the procedure that led from (B2) to (B4), we get

$$(B6) \quad \partial \text{Cov}[y, A^{\nu-1}(y)] / \partial p_1 = \text{Cov}\{x_1, [1 - F(y)]^{\nu-1}\}.$$

Now that the derivative of the numerator is known, we can derive the derivative of the extended Gini with respect to p_1 .

$$(B7) \quad \partial G_y(\nu) / \partial p_1 = S_1 c(x_1, y, \nu) - S_1 G_y(\nu)$$

where $S_1 = m_1 / m_y$ is the share of x_1 and $c(x_1, y, \nu) = -\nu \text{Cov}\{x_1, [1 - F(y)]^{\nu-1}\} / m_1$ is the (weighted) concentration index.

The effect of an equal-revenue tax change on the extended Gini can be evaluated by using (B7). Let dG/dR denote the derivative of the extended Gini with respect to a tax reform; that is,

$$(B8) \quad dG/dR = [\partial G_y(\nu)/\partial p_1] dp_1 + [\partial G_y(\nu)/\partial p_2] dp_2.$$

By using the equal-yield revenue restriction, $m_1 \alpha_1 dp_1 + m_2 \alpha_2 dp_2 = 0$, and (B7) we get

$$(B9) \quad dG/dR = [c(x_1, y, \nu) - \alpha_{12} c(x_2, y, \nu)] S_1 dp_1$$

which means that the effect of the tax reform on the extended Gini depends on the difference between the concentration ratios of the commodities and the deadweight loss of the tax system. Note that we assumed $\alpha_{12} = 1$; that is, the change in the deadweight loss is zero. This completes the proof of property (a).

PROOF OF PROPERTY (b):

Let x represent the expenditure on a commodity; then, the (extended Gini) concentration ratio is defined as

$$(B10) \quad c(x, y, \nu) = -\nu \text{Cov}\{x, [1 - F(y)]^{\nu-1}\} / m_x.$$

Writing the covariance explicitly and eliminating zeros yields

$$(B11) \quad c(x, y, \nu) = -(\nu/m_x) E_y E_x \{ (x - m_x) [1 - F(y)]^{\nu-1} \}.$$

Now let $g(y) = E_x(x|y)$, so that $g(y)$ is the Engel curve. Then (B11) can be written as

$$(B12) \quad c(x, y, \nu) = (-\nu/m_x) \times \int_0^\infty [g(y) - m_x] [1 - F(y)]^{\nu-1} f(y) dy$$

and by integration by parts, where

$$V(y) = [1 - F(y)]^{\nu-1} \\ U'(y) = [g(y) - m_x] f(y)$$

and by rearranging terms, we get

$$(B13) \quad c(x, y, \nu) = [-\nu(\nu-1)/m_x] \\ \times \int_0^\infty \int_0^y [g(t) - m_x] f(t) dt \\ \times [1 - F(y)]^{\nu-2} f(y) dy.$$

By transformation of variables, where $F = F(y)$ and

$$\Theta_x(F) = (1/m_x) \int_0^F g(y(F)) dF$$

is the concentration curve, we get

$$(B14) \quad c(x, y, \nu) = -\nu(\nu-1) \\ \times \int_0^1 [\Theta_x(F) - F] [1 - F]^{\nu-2} dF$$

which means that in the case of the Gini ($\nu = 2$) the concentration ratio is equal to twice the area between the concentration curve and the 45° line. In other cases, the (extended) Gini concentration ratio is equal to a (weighted) difference of the area between the 45° line and the concentration curve. Note that the difference between concentration ratios is hence equal to the (weighted) difference between concentration curves.

REFERENCES

- Atkinson, Anthony B., "On the Measurement of Inequality," *Journal of Economic Theory*, September 1970, 2, 244-63.
Calmus, Thomas W., "Measuring the Regressivity of Gambling Taxes," *National Tax Journal*, June 1981, 34, 267-70.

- Clotfelter, Charles T., "On the Regressivity of State-Operated 'Numbers' Games," *National Tax Journal*, December 1979, 32, 543-8.
- Diewert, W. Erwin, "Optimal Tax Perturbations," *Journal of Public Economics*, October 1978, 10, 139-77.
- Dixit, Avinash K., "Welfare Effects of Tax and Price Changes," *Journal of Public Economics*, February 1975, 4, 103-23.
- _____, and Munk, K. J., "Welfare Effects of Tax and Price Changes: A Correction," *Journal of Public Economics*, August 1977, 8, 103-7.
- Formby, John P. and Sykes, David, "State Income Tax Progressivity," *Public Finance Quarterly*, April 1984, 12, 153-65.
- _____, Smith, W. James and Sykes, David, (1986a) "Intersecting Tax Concentration Curves and the Measurement of Tax Progressivity," *National Tax Journal*, March 1986, 39, 115-21.
- _____, _____ and _____, (1986b) "Income Redistribution and Local Tax Progressivity: A Reconsideration," *Canadian Journal of Economics*, November 1986, 19, 808-11.
- Goldie, Charles M., "Convergence Theorems for Empirical Lorenz Curves and Their Inverses," *Advances in Applied Probability*, December 1977, 9, 765-91.
- Guesnerie, Roger, "On The Direction of Tax Reform," *Journal of Public Economics*, April 1977, 7, 179-202.
- Hadar, Josef and Russell, William R., "Rules for Ordering Uncertain Prospects," *American Economic Review*, March 1969, 59, 25-34.
- Hanoch, Giora and Levy, Haim, "The Efficiency Analysis of Choices Involving Risk," *Review of Economic Studies*, July 1969, 36, 335-46.
- Hatta, Tatsuo and Haltiwanger, John, "Tax Reform and Strong Substitutes," *International Economic Review*, June 1986, 27, 303-15.
- Kakwani, Nanak C., (1977a) "Applications of Lorenz Curves in Economic Analysis," *Econometrica*, April 1977, 45, 719-27.
- _____, (1977b) "Measurement of Tax Progressivity: An International Comparison," *The Economic Journal*, March 1977, 87, 71-80.
- _____, *Income Inequality and Poverty*, New York, Oxford University Press, 1980.
- Kiefer, Donald W., "Distributional Tax Progressivity Indexes," *National Tax Journal*, December 1984, 37, 497-513.
- Latham, Roger, "Lorenz-Dominating Income Tax Function," *International Economic Review*, February 1988, 29, 185-200.
- Lerman, Robert and Yitzhaki, Shlomo, "Improving the Accuracy of Estimates of Gini Coefficients," *Journal of Econometrics*, September 1989, 42, 43-7.
- Liu, Pak-Wai, "Lorenz Domination and Global Tax Progressivity," *Canadian Journal of Economics*, May 1985, 18, 395-9.
- Mayshar, Joram, "A Note on Measuring the Marginal Cost of Taxation," Working Paper No. 175, Department of Economics, Hebrew University, Jerusalem, 1988.
- Mera, Koichi, "Empirical Determination of Relative Marginal Utilities," *Quarterly Journal of Economics*, August 1969, 83, 464-77.
- Musgrave, Richard, A., *Theory of Public Finance*, New York: McGraw-Hill, 1959.
- Pfähler, Wilhelm, "Redistributive Effects of Tax Progressivity: Evaluating a General Class of Aggregate Measures," *Public Finance*, 1987, 42 (1), 1-31.
- Piggott, John, "The Social Marginal Valuation of Income: Australian Estimates from Government Behaviour," *The Economic Record*, March 1982, 58, 92-9.
- Shorrocks, Anthony F., "Ranking Income Distributions," *Economica*, February 1983, 50, 3-17.
- _____, and Foster, J. E., "Transfer Sensitivity Inequality Measures," *Review of Economic Studies*, July 1987, 54, 485-97.
- Suits, Daniel B., (1977a) "Gambling Taxes: Regressivity and Revenue Potential," *National Tax Journal*, March 1977, 30, 19-35.
- _____, (1977b) "Measurement of Tax Progressivity," *American Economic Review*, September 1977, 67, 747-52.
- Weisbrod, Burton A., "Income Redistribution Effects and Benefit-Cost Analysis," in S. B. Chase, ed., *Problems in Public Expenditure Analysis*, Washington, DC: The Brookings Institution, 1968, pp. 177-209.

- Whitmore, G. A., "A Third-Degree Stochastic Dominance," *American Economic Review*, June 1970, 60, 457-9.
- Wibaut, Serge, "A Model of Tax Reform in Belgium," *Journal of Public Economics*, February 1987, 32, 53-77.
- Wildasin, David E., "On Public Good Provision with Distortionary Taxes," *Economic Inquiry*, April 1984, 22, 227-43.
- Yaari, Menahem E., "A Controversial Proposal Concerning Inequality Measurement," *Journal of Economic Theory*, April 1988, 44, 381-97.
- Yitzhaki, Shlomo, "Stochastic Dominance, Mean Variance, and Gini's Mean Difference," *American Economic Review*, March 1982, 72, 178-85.
- _____, "On An Extension of the Gini Inequality Index," *International Economic Review*, October 1983, 24, 617-28.
- _____, "On the Progressivity of Commodity Taxation," Discussion Paper No. 255, Development Research Department, The World Bank, 1987.
- _____, and Olkin, Ingram, "Concentration Curves," Working Paper No. 179, Department of Economics, The Hebrew University, 1988.
- _____, and Thirsk, Wayne, "Welfare Dominance and the Design of Excise Taxes in Côte d'Ivoire," *Journal of Development Economics*, July 1990, 33, 1-18.

The Role of Demandable Debt in Structuring Optimal Banking Arrangements

By CHARLES W. CALOMIRIS AND CHARLES M. KAHN*

Demandable-debt finance by banks warrants explanation because it entails costs of bank suspension, liquidation, and idle reserve holdings. An explanation is developed in which demandable debt provides incentive-compatible intermediation where the banker has comparative advantage in allocating investment funds but may act against the interests of uninformed depositors. Demandable debt attracts funds by giving depositors an option to force liquidation. Its usefulness in transacting follows from information-sharing between monitors and nonmonitors. (JEL G21)

For centuries, the vast majority of externally financed investments have been funded by banks, for which demandable-debt instruments (bank notes and checking accounts) have been the principal source of funds. The goal of this paper is to explain the emergence of demandable-debt banking historically as the primary means of external finance in the economy.

Demandable debt warrants explanation because, in several respects, it appears more costly than available alternative contracting structures. By issuing demandable debt, banks created a mismatch between the maturity of assets and liabilities. This mismatch left them exposed to the possibility that depositors would attempt to withdraw more

funds than a bank could supply on short notice. When this occurred, the consequences were costly. Individual banks that did not meet their obligations were forced into expensive procedures (liquidation or receivership) that would not have arisen in an equity-based or maturity-matched contracting structure.¹ If depositors en masse attempted to withdraw funds from the entire banking system, banks as a group were forced to suspend convertibility of their liabilities into specie on demand. Such suspension was also disruptive and costly. To defend against either of these undesirable consequences, banks had to hold a proportion of their assets in idle reserves to insulate themselves from excessive withdrawals.

Given these costs, demandable debt seems inferior to both maturity-matched debt and equity contracting. However, in this paper, we show that demandable debt has an important advantage as part of an incentive scheme for disciplining the banker. In effect, demandable debt permits depositors to "vote with their feet"; withdrawal of funds is a vote of no-confidence in the activities of the banker. Without the ability to make early withdrawals, depositors would have little incentive to monitor the bank.

*Department of Economics, Northwestern University, Evanston, IL 60208, and Department of Economics, University of Illinois, Urbana, IL 61801, respectively. We thank Lee Alston, Herbert Baer, Kyle Bagwell, Ben Bernanke, Sudipto Bhattacharya, Doug Diamond, Gary Gorton, Monica Hargraves, Charlie Jacklin, Dick Jefferis, and participants in the joint Northwestern-University of Chicago theory seminar and seminars at the Federal Reserve Bank of Chicago, Purdue University, SUNY Stony Brook, the University of Illinois, the National Bureau of Economic Research, and the Garn Institute for helpful comments. The initial work was partially funded by the National Science Foundation under grant SES-8511137. We are grateful to the Garn Institute of Finance and the Herbert V. Prochnow Educational Foundation of the Graduate School of Banking, Madison, WI, for additional support.

¹Kenneth R. Cone (1983) shows that, in a world of full information, the risk of depositor liquidation under demandable debt is absent, provided that financial intermediaries are maturity-matched.

This account gives a natural rationale for two important institutional features of banking. The so-called "sequential service constraint," by which payments were made to demanders on a first-come, first-served basis, becomes intelligible as a way to make monitoring depositors interested in registering their no-confidence votes at the first opportunity. The ease with which banks may be forced into liquidation, far from being an unfortunate consequence of the contracting structure, turns out to be central to the structure: we show that, by submitting to the threat of liquidation under appropriate circumstances, the banker can reduce his cost of capital.

In addition, our account may have wider applicability. Features of modern capital structures of nonfinancial institutions bear important similarities to the historical role of demandable debt. Modern-day firms often have multilayered debt structures, in which certain debt-holders have priority of claim for repayment. Claimants to short-term senior debt in modern firms may play a similar role to that of the monitoring depositors in our model.

The paper is organized as follows: In Section I, we contrast our explanation of demandable debt with the literature based on desire for flexibility of consumption. The model in Section II demonstrates the value of a demandable-debt contract in the case of a single investor contracting with the banker monopolist. Here, a run corresponds to a demand by the investor for liquidation of the bank. Section III examines the case in which different monitors receive different (independent and identically distributed) signals. In this case, it pays to have more than one depositor monitoring the bank, because the quality of signals in the aggregate improves with the number of monitors. Banks find it advantageous to hold reserves to provide a buffer that reduces the likelihood of unwarranted liquidation. An optimal threshold of withdrawal orders is chosen at which the bank is liquidated, and relative payoffs ensure that the optimal number of monitors invest in receiving signals.

At the end of Section III, we briefly and informally indicate how solving the incentive problem facing the banker will also make the banker's liabilities more transactable. Formal models combining the incentive problem and liquidity are an important field for further research.² Section IV summarizes and indicates important limitations of our results.

I. Explanations for Demandable Debt

Recent theoretical work on the role of banks has tended to divide into two categories. Theory in one category emphasizes the role of banks as providing flexibility for depositors in the timing of consumption. Theory in the other category, to which our paper belongs, emphasizes the *incentive* problem inherent in the divergence of interest between a bank's depositors and its managers.³ For reasons indicated below, we believe that accounts which ignore the incentive problem facing the banker do not adequately explain why banks historically settled on demandable debt.

A. Consumption Flexibility and Demandable Debt

In the past several years, the preeminent theoretical analyses of banks, bank runs, and bank regulation have assumed that the economic role of demandable debt is to provide flexibility to risk-averse depositors who are uncertain about the timing of their future consumption demand.⁴ In this category of models, bank runs, when they occur, are an unfortunate and undesirable side-

²See Gary Gorton and George Pennacchi (1988), Charles J. Jacklin (1988), and A. P. Villamil (1988) for various approaches to combining the liquidity and incentive arguments.

³Jacklin and Sudipto Bhattacharya (1988) give a concise but useful review of these approaches.

⁴Fundamental papers that utilize this approach are by John Bryant (1980), Douglas W. Diamond and Philip Dybvig (1983), and Jacklin (1987). For a model emphasizing the costs to depositors of delay in liquidation, see Merwan Engineer (1987).

effect of a contract whose whole purpose is to provide consumption flexibility.

Although these models provide both a concise formalization of the fact that banks provide consumption flexibility and a coherent account of bank runs, they are unable to account for several important institutional features of demandable debt. First, in the absence of incentive constraints on the part of the banker, the optimal arrangement in liquidity-based accounts always involves suspension of convertibility, rather than expensive liquidation. However, suspension was not an option available to individual banks; it was only an alternative for the financial system as a whole, in the face of system-wide panics. Individual banks that could not satisfy creditors' fears about solvency were not permitted to suspend; they were forced to close.⁵

Second, studies of actual bank failures give fraud a prominent place in the list of causes. Studies of 19th- and 20th-century banking indicate that fraud and conflicts of interest characterize the vast majority of bank failures for state and nationally chartered banks.⁶

⁵See Calomiris and Larry Schweikart (1988) for a discussion of suspension rules during the early U.S. experience. Kevin Dowd (1988) argues that individual-bank suspension of debt redemption would have been beneficial but was prevented by legal prohibitions. We argue that the prohibition-of-suspension option clauses simply reflected the learned desirability of placing the decision regarding whether suspension was "justified" outside the control of the individual banker. The legal prohibition of option clauses on notes may have been perceived as necessary to protect some unsophisticated note-holders, while no such law was deemed necessary for relatively sophisticated depositors.

⁶For example, E. L. Smead (1928) found that three of the nine most common causes of bank failure in the 1920's involved fraudulent or questionable activities by the banker: loans to officers and directors, outright defalcation, and loans to enterprises in which officers and directors were interested. For discussions of the role of fraud in earlier eras, see Carter H. Golembe and Clark Warburton (1958), George J. Bentson and George G. Kaufman (1986), and Calomiris and Schweikart (1988). Data on national bank failures, by cause, can be found in the Annual Report of the Comptroller of the Currency (1920 pp. 56-79). For information on the importance of fraud in more recent bank failures, see Comptroller of the Currency (1988).

Third, receivership resulted from a critical mass of depositor withdrawal orders and was invoked because of information about bank asset values, not because of exogenous liquidity needs of individual depositors. In cases of massive exogenous demand for an individual bank's assets by small depositors, banks avoided failure by appealing to other banks for loans of reserves; however, when large informed depositors (including other bankers) concluded that a bank was in trouble, they would precipitate a run, depleting the bank's reserves and forcing it to be placed in receivership.⁷

These considerations make it apparent that the liquidation of banks—which was part and parcel of demandable-debt contracts—was designed to place the assets of banks beyond the reach of the banker. The rationale for prohibiting banks from suspending at their own discretion may have been the discipline that it imposed on the behavior of the banker. Thus, a model of demandable debt with bank liquidation through receivership should account for the desirability of taking control of the bank away from the banker at the option of depositors.

Fourth, the "sequential-service constraint" (first-come, first-served rule) for bank withdrawals, which allowed informed depositors to receive repayment before banks were placed into receivership, also

⁷Henry C. Nicholas (1907 p. 26) dismissed the importance of withdrawals by uninformed depositors in causing bank liquidation. He wrote, "If a bank is actually in bad shape there is far more likelihood of its initial condition being discovered by other banking institutions than by the individual depositors of the bank A run is sometimes started in this manner ... and continues until it has practically wiped out the reserves of the suspected institution, the ordinary depositors receiving their first information regarding the position of the bank when that institution is finally forced to close its doors and formally apply for a receiver." This discussion makes important points about bank runs which appear in our model: some depositors are informed, while others are not. Runs by informed depositors end in liquidation. Informed depositors are able to exercise their withdrawal option before uninformed depositors are able to observe the bank's difficulty (or the run).

warrants explanation. In cases other than banking, payments from bankrupt firms to creditors in anticipation of bankruptcy are not allowed, and creditors may be forced to relinquish such payments during the bankruptcy proceeding. Why in the case of banking should those who run the bank receive preferential treatment in liquidation states?

B. Demandable Debt as an Incentive Scheme

Models in the second category of theory on the role of banks begin with the assumption that bankers have an informational advantage in determining which projects are most worthy of financing. Therefore, the banker has a comparative advantage in allocating funds for investment, but he also may have the ability to act against the interests of uninformed depositors.⁸

We show that demandable debt can provide an incentive-compatible solution to this problem in the presence of costly information. The right to take one's money out of the bank if one becomes suspicious that realized returns are low makes it in the depositor's interest to keep an eye on the bank. If enough depositors agree with this negative assessment of the bank's future, liquidation will be called for, and the bank

will close. The demandable-debt contract allows the banker to precommit to a set of payoffs he otherwise would not be able to offer depositors.

Not all depositors need monitor the banker. We argue that the first-come, first-served (sequential service) rule of demandable debt provides compensation for those who choose to invest in information and thus avoids free-riding. We view bank intermediation, therefore, as a three-sided relationship. The monitors pay the costs of vigilance but receive the benefit of knowing that they will be "first in line" (and thereby receive a higher payment than other depositors) should it become necessary to withdraw their funds from the bank. The depositors who do not monitor are willing to pay the price of being last in line in "bad" states, because they receive a benefit in return: the active monitors keep the banker in line and thereby provide a benefit to the passive depositors. Depositors need not reveal whether they are active or passive; the same contract works for both types.

The physical structure we assume includes the following important features. 1) The bank is operated by a monopolist with special access to a profitable investment opportunity which yields either a good or a bad realization. 2) There is potential for cheating by the banker which takes the form of his absconding with a proportion of the bank's assets after the investment realization. (One can think of this more generally as costly *ex post* fraudulent behavior which the banker undertakes whenever it is more profitable to do so than to make the promised payments to depositors.) 3) Depositors face different costs of obtaining a signal that allows them to predict profitability. 4) An authority exists who will enforce contracts (some of which may stipulate conditions for bank liquidation) and who can act as receiver for liquidated banks. 5) Depositors have a reservation level of return on their endowments below which they will not invest funds with the banker.

The profit-maximizing banker will act to maximize social gain by selecting a contract that achieves beneficial intermediation (investment in profitable enterprises), while

⁸This point is emphasized by Diamond (1984) and Ben Bernanke and Mark Gertler (1987). For an overview of the relation between agency costs and the structure of financial contracts, see Eugene F. Fama (1988). Diamond's solution to the delegated-monitoring problem of financial intermediation relies on two assumptions that are absent in our framework: the existence of an *ex post* nonpecuniary penalty that can be imposed on the banker and the ability of the banker to construct a riskless portfolio through diversification. The second assumption permits enforcement of the penalty, even if cheating is costly to observe directly, whenever the banker fails to meet his obligations. Bernanke and Gertler provide a simple macroeconomic model in which bankers are subject to moral hazard and depositors desire liquidity. They explicitly assume that costly monitoring and punishment of defaulting bankers are not possible. For them, demandable debt is desirable solely for its liquidity. In our model, demandable debt is desirable although liquidity demand is absent.

avoiding as much as possible the costs associated with absconding or liquidating. We find that the demandable-debt contract is optimal for a range of parameter values. The potential for costly liquidation may be more than offset by the social gain that comes from enhanced investment opportunities.⁹

II. The Model with a Single Depositor

A. Physical Structure

A banker has an investment opportunity, but he lacks sufficient capital to take advantage of it. The investment opportunity costs one dollar. Each potential depositor has one dollar to invest. We will let S represent the total expected return available for a dollar's investment elsewhere in the economy. We assume that all agents are risk-neutral; thus, any scheme the banker develops will have to yield a depositor that same expected return.

The investment opportunity yields an uncertain payoff which may take one of two values, T_1 or T_2 , with $T_2 > T_1$. The probabil-

ity of the high outcome is γ . The realization is unknown to all parties at the outset and is observable *ex post* only by the banker. Thus, there is no way to make a contract tied directly to the value of T_i .¹⁰

Let period 3 be the date at which the payoff is realized and the loan is to be repaid. We assume that in period 3, immediately before repayment, the banker has the opportunity to abscond with the funds. Absconding is socially wasteful; for concreteness, we will assume that it reduces the realization T_i by the proportion A , where A is between 0 and 1.

Although the act of absconding reduces the size of the pie that is divided between the banker and the depositor, it places the banker beyond the reach of the law. Therefore, he is no longer constrained to repay the loan as initially promised. Thus, any promise to pay the depositor an amount P is actually an option of the banker either to pay P or to leave town with his assets diminished by the proportion A .

The losses from absconding may be interpreted in a variety of ways. They may represent the cost of engaging in fraud (payments to coconspirators) or the costs (forgone earnings) of placing the bank's resources in a form that allows theft. The latter interpretation requires a richer, multiperiod model than the one we provide, in which bankers' allocation decisions depend on last-period earnings.¹¹

It should be readily apparent that the temptation to abscond will be greater with lower realizations of T_i . In deciding whether

⁹V. Chari and Ravi Jagannathan (1988) provide an example of an information-based run for a model that has many features in common with ours. A key difference is that they assume an (exogeneously imposed) negative externality from liquidation of the bank's assets. In their model, the creation of a liquidation technology is not efficient. In our model, there is a positive externality from running the bank: when the depositor observes a bad signal, he calls for liquidation, thereby salvaging some of the bank's value. The bank's structure is designed to internalize this positive externality and allows nonmonitoring depositors to compensate monitors for the benefits they provide.

Our model can also be interpreted as allowing depositors to exercise a put option based on the information they receive. However, unlike the usual "inside-trading" scenario, the uninformed depositors also benefit at the expense of the bank. While the uninformed depositors receive a lower payoff than the informed depositors, they benefit because the bank is prevented from cheating. In the usual scenario (e.g., Albert S. Kyle, 1981), the uninformed either lose or the informed cannot successfully earn a return on their information-production because of free-riding, as in Sanford J. Grossman and Joseph E. Stiglitz (1980). We thank an anonymous referee for suggesting this comparison to us.

¹⁰We assume that the banker is not able to trade in equity shares. This conforms with the relative illiquidity of equity trade in the period under examination. It could also be generated as a conclusion in a model in which bankers possess specialized information about investment projects of borrowers. Robert M. Townsend (1979) notes that in circumstances when only one party has access to information, debt contracts (i.e., contracts not contingent on the private information) will often be the only feasible alternative.

¹¹The plausibility of our "leaky bucket" assumption and possible multiperiod reinterpretations are discussed further in the final section of the paper. For an initial generalization of the absconding assumption see Calomiris et al. (1990).

to abscond, the banker compares the "tax" on absconding, AT_1 , with the promised funds due the depositor. If the absconding tax is less, then absconding is more profitable than paying up. Historical evidence confirms the greater prevalence of fraud in times of low returns to bank investments.¹²

Because of the threat that the banker will abscond—a threat against which he cannot commit himself—it will generally be necessary for the banker to increase the payment offered to a depositor by a "default premium" as protection against those states in which the depositor will, in fact, receive nothing.

Note that the addition of a default premium can, in turn, increase the probability of default, by making it desirable for the banker to abscond in good states as well. For example, suppose

$$S > AT_1$$

so that any payment promised to the depositor must be sufficiently large to incur absconding in the low realization; that is, a promise to pay P will only be honored a fraction γ of the time. Suppose also that

$$\gamma T_2 + (1 - \gamma)(1 - A)T_1 > S$$

so that the investment would be socially desirable (even taking into account the loss from absconding in the low realization). Then, if

$$S > \gamma AT_2$$

there is no way to promise the depositor enough expected payment to make him willing to invest, despite the social desirability of the project; the promised payment would have to exceed AT_2 , making it desirable for the banker to abscond all the time.

¹²The concentration of bank fraud during times of regional or national economic decline is pronounced in national bank-failure data. See the Annual Report of the U.S. Comptroller of the Currency (1920 pp. 56–79).

Because of the loss of socially desirable opportunities, it is useful to have a method of thwarting absconding. One such method is the liquidation of the bank in period 2. Liquidation means that the bank's assets are taken over by a receiver, controlled by a court. This is an expensive process, not the least because the court-appointed and court-controlled receiver is likely to be less able to realize the full potential of the assets. On the other hand, the fact that the assets are no longer in the banker's control preempts any decision by him to abscond with the funds.

We assume that liquidation reduces the value of the assets by the proportion L , so that L can be regarded as the tax due to liquidation. For a complete characterization of the process of liquidation, it is necessary to take some stand as to the maximum that can be feasibly paid to the depositor in the case of liquidation. We call this value M , and we assume that¹³

$$(1) \quad AT_2 > M > AT_1$$

so that the amount that can be guaranteed to the depositor in a liquidating contract is greater than the maximum amount that can be guaranteed in a nonliquidating contract. We also assume that

$$(2) \quad L > A$$

so that liquidation is less wasteful socially than is absconding.¹⁴

¹³There are several ways we can approach the question of the maximum to be paid once the court has control. For simplicity, we assume that M does not vary with the realization of T_1 . One argument is that the value of the firm might be determined by the court, but at a very high cost.

¹⁴Actual liquidation costs in the United States varied historically, depending on time, location, and bank size but seem to have been small relative to potential social losses from absconding, as our model assumes. Bankruptcy expenses averaged between three percent and six percent of total collections for national banks between 1872 and 1904 (Brian C. Gendreau and Scott S. Prince, 1986).

In some cases, it may be desirable *always* to put the assets of the bank into liquidation rather than risk the banker's absconding. We call such an agreement a "simple liquidation contract," as opposed to a "simple nonliquidation contract," which states a promised repayment and leaves it to the banker whether to abscond or not.

The more interesting case, however, is one in which the depositor, based on his own information, is given the option of demanding liquidation or not. Specifically, suppose that by paying a cost I the depositor is able to receive a signal σ in period 1 as to the likelihood of a high (T_2) or low (T_1) realization. The action of investing in the signal and the result of this action are private. The signal σ works as follows. It takes on one of two values $\{g, b\}$ (for "good" and "bad").¹⁵ The probability of a high realization, contingent on the signal, is ρ_σ :

$$(3) \quad \rho_g > \gamma > \rho_b.$$

We will use the indicator variable $e \in \{0, 1\}$ to represent the depositor's choice: $e = 1$ if there was an investment in the signal, 0 otherwise.

In summary, the physical structure of our model is as follows. There are three periods. In period 1, the depositor may invest in receiving a signal. In period 2, the bank may be liquidated. In period 3, the loan is repaid to the depositor, unless the banker decides to abscond (which he can only do if the bank has not been liquidated).

B. The Contracting Structure

Contracts are arranged in period 0. The monopolist banker offers the profit-maximizing contract among those which yield the depositor at least S in expected returns. (If no such contract exists or the best such contract yields negative profits, then none is offered.)

¹⁵In the single-depositor case, the assumption that the signal takes only two values is not restrictive. In fact, the multidepositor model of the subsequent section can be reinterpreted as a single-depositor model with multivalued signals.

The universe of contracts in this structure is as follows. A contract is a function from a space of announcements Σ into *outcomes*. An outcome is a pair (P, Λ) , where $\Lambda \in \{0, 1\}$ is an indicator variable equaling 1 if liquidation is mandated and 0 otherwise. P is the mandated repayment. (Of course P will only be received if the banker does not abscond.)¹⁶

If the contract only specifies one outcome, we call it a "simple contract"; otherwise we call it a "compound contract." We have already described the two kinds of simple contracts: the simple liquidating contract and the simple nonliquidating contract. A straightforward application of the revelation principle demonstrates that, for the single depositor case, contracts need never contain more than two outcomes, because the signal the depositor may observe has only two values. We can identify the announcements in a compound contract with assertions by the depositor that he has observed one or the other signal. Thus, a compound contract consists of a quartet $(P_b, \Lambda_b, P_g, \Lambda_g)$.

Each contract generates a sequential game in which the depositor chooses the level of investment in information-gathering (e) and the announcement he makes as a function of the signal he receives. The banker chooses whether to abscond as a function of the announcement made by the depositor and the realization on the investment. An *optimal* contract is one for which there is a sequential equilibrium that generates maximum profits consistent with the depositor's receiving expected returns equal to the amount S .

¹⁶As it stands, the specification of the contract is incomplete in two technical respects. First, the specification of the outcome should include a specification of the banker's response (i.e., whether he chooses to abscond) as a function of the announcement $\hat{\sigma}$ and of the realization T_i . However, in almost all contracts, the banker's response is easily discerned: he absconds if $P_{\hat{\sigma}} > gAT_i$ and does not abscond if $P_{\hat{\sigma}} < AT_i$. Only in the case of indifference would it be necessary to specify his response in detail. Second, the contract does not include the possibility of randomized outcomes. These can be shown never to dominate deterministic outcomes.

THEOREM 1:¹⁷ *The optimal contract in the problem takes one of the following four forms:*

- a) *a simple nonliquidating contract*
- b) *a simple liquidating contract; in this case,*

$$AT_1 < P \leq M$$

- c) *a compound contract composed of two simple nonliquidating contracts ($\Lambda_b = \Lambda_g = 0$); in this case,*

$$P_b \leq AT_1 \text{ and } AT_1 < P_g \leq AT_2$$

- d) *a compound contract composed of one simple liquidating contract and one simple nonliquidating contract ($\Lambda_b = 1, \Lambda_g = 0$); in this case,*

$$AT_1 < P_b < P_g \leq AT_2.$$

If the optimal contract is a compound contract, then the depositor invests in the signal; if it is a simple contract, he does not. In the case of compound contracts, absconding occurs if and only if the signal was *g* but the low-value outcome T_1 was realized.

We call contract d "demandable debt." It works as follows: after making the deposit, the depositor invests in learning what the likely outcome will be. If he receives the bad signal, he opts for liquidating the bank. This delivers a payment with certainty. If he receives the good signal, he opts for not liquidating the bank. This promises a higher payment but runs the risk of the banker's absconding.

Contract c works in virtually the same way. The only difference is that the guaranteed payment in the case of a bad signal is sufficiently low that the banker will never wish to abscond and so it is not necessary to use liquidation to hold him in place. Since liquidation always involves social costs, it is not difficult to demonstrate that in any case where contract c is feasible, it dominates contract d. We will (with prejudice) describe contract c as a "nuisance contract."

Next, we provide a characterization of when the various contracts will be observed.

We do so under the assumption that the signal is "accurate" (i.e., ρ_g is high and ρ_b is low, so that the signal is a good predictor of the state) and the signal is "cheap" (so that I is small). It is easily demonstrated that, if the signal is sufficiently inaccurate or sufficiently expensive, a compound contract is not useful.

THEOREM 2: *If the signal is sufficiently cheap and accurate; then there exist values S^* and \hat{S} , such that the optimal contract depends on the required returns S in the following way: for $S \leq AT_1$ the simple, nonliquidating contract is optimal; for $S \in (AT_1, S^*]$, the nuisance contract is optimal; for $S \in (S^*, \hat{S}]$, demandable debt is optimal; and for $S > \hat{S}$, no contract is feasible.*

In other words, demandable debt will be observed when the returns that depositors can receive in alternate investments are relatively high.

III. Multiple Depositors with Independent Signals

In this section, we develop a model for the case in which a number of depositors enter into contracts with the banker. As before, each depositor has one dollar to invest, and the banker has one "project" he can pursue. The project costs Y and yields a total return of YT_i , which takes one of two values. Any deposits the banker receives in excess of Y can be used to yield the same competitive return S that depositors have available to them on their own. Deposits in excess of Y will be identified with "reserves."

We make the following natural assumptions about the difference between the two forms of bank assets, "project" and "reserves." If the bank is liquidated, the value of the project decreases by $1 - L$; the value of the reserves is unchanged.¹⁸ If the banker

¹⁷Proofs of theorems are outlined in the Appendix.

¹⁸This assumption is natural, given that we regard the project as requiring the banker's expertise and regard the reserves as invested in publicly available technologies.

TABLE 1—PAYOFFS ON EACH OF THE THREE NODES OF THE GAME TREE

Contract	Banker receives	Depositors receive
Liquidation	$(1-L)T_1Y + (Z-Y)S - P$	P
No liquidation		
Banker absconds	$(1-A)T_1Y$	$(Z-Y)S$
Banker does not abscond	$T_1Y + (Z-Y)S - P$	P

absconds, then he takes the projects with him and receives $(1-A)YT_1$. The depositors retain the entirety of the reserves.¹⁹ We strengthen assumption (2) as follows:

$$(4) \quad L < A(T_1/T_2).$$

There are Z individuals available to enter into a contract with the bank. Of these individuals, K can receive signals by investing at a cost I ; for the remainder, the cost of receiving a signal is prohibitive.²⁰ Signals are independent and identically distributed (i.i.d.) conditional on T_i . For any individual, a "bad" signal is associated with reduced likelihood of the high-productivity state T_2 , so $\rho_b < \rho_g$, as before.

Supposing that all K individuals have invested in the signal, let N be the number who receive the "bad" realization. Given the i.i.d. structure, N is a sufficient statistic for T_i , and the probability that the realization is T_2 decreases with N .

A. The Contract from the Banker's Viewpoint

We start by examining only the incentive problem for the banker, taking the behavior of all depositors as given. We will return to

the individual depositors' incentives in the succeeding subsection. For now, we assume that all K individuals who can invest in obtaining the information do so and report it truthfully.²¹ A *contract* specifies an aggregate payment P and a liquidation decision Λ as functions of the number of depositors who announce observations of the bad signal. (In the succeeding subsection, we will investigate a scheme for dividing aggregate payments among the depositors.) Note therefore that the contract is the direct generalization of the contract in the previous section to a case of multiple signals.

After the announcement of the signals, the game tree is as before: if a liquidation is not mandated, the banker makes a decision whether to abscond. Table 1 describes the payoffs on each of the three nodes of the game tree.

The *optimal contract* maximizes the banker's expected profits subject to three restrictions.

- 1) The expected payments to the depositors equal their aggregate reservation level:

$$SZ + KI.$$

That is, all depositors must be compensated for the opportunity cost of their funds; in addition, any monitors must be compensated for the cost of monitoring.

- 2) In the case of liquidation, actual payment cannot exceed what is assumed feasible; as before, we suppose that a liquidated investment Y pays off at most MY to the depositors. Thus, the total pay-

¹⁹An alternative assumption is that, if the banker absconds, he takes the entirety of the reserves as well. The assumption in the text is natural if we regard absconding as occurring by siphoning a project into a less desirable project whose returns accrue directly to the banker. The assumption in this footnote is natural if we regard absconding as occurring when the banker piles the loot into the stagecoach and heads out of town.

²⁰This is the simplest structure of supply of signals; it can be generalized. Alternatively, the cost of investing in a signal could be determined in a general equilibrium model.

²¹It will be clear that, as long as the cost of investing in the signal is sufficiently low, it is optimal to have all individuals with cost I make the investment.

ment to depositors out of the project and the reserves is

$$P \leq MY + (Z - Y)S \quad \text{if } \Lambda = 1.$$

- 3) Finally we must consider the banker's incentive to abscond. If liquidation does not occur, then the banker will prefer to abscond whenever

$$AT_1Y < P - (Z - Y)S.$$

If the inequality is reversed the banker prefers not to abscond.

As before, we define \hat{S} to be the least upper bound of feasible expected returns to depositors from the project; if the required rate of return exceeds \hat{S} , no contract is feasible. \hat{S} can be calculated explicitly.

Our first result is that, for required returns which are sufficiently high (but less than \hat{S}), the optimal contract calls for liquidation when the number of bad signals is high, and not when the number of bad signals is low. When the number of bad signals is low, there is a positive (but small) probability that the banker will abscond.

THEOREM 3: *For an interval of values of S , $(\underline{S}, \hat{S}]$, the optimal contract has the following form: there exists \underline{N} such that:*

If $N > \underline{N}$, $\Lambda(N) = 1$ and

$$P(N) = MY + (Z - Y)S;$$

If $N < \underline{N}$, $\Lambda(N) = 0$ and

$$P(N) = AT_2Y + (Z - Y)S.$$

In other words, the contract has informed agents announce whether their signal was bad. If more than a critical number \underline{N} announce bad signals, the bank is liquidated. If fewer than \underline{N} announce bad signals, the bank is not liquidated, and the banker chooses to abscond if the productivity draw was low.²²

²²If exactly \underline{N} announce bad signals, the optimal contract has a randomization between liquidation and nonliquidation. We omit the details.

Note that Z is arbitrary in this contract. As Z increases, the optimal P increases one-for-one: additional deposits beyond those invested in the project are held in reserves and returned to the depositors with certainty.²³

B. Depositor Incentives

It remains to be shown that the total aggregate payment to depositors specified in the previous section can be divided among depositors in such a way as to maintain the incentives for low-cost-information depositors to invest in the signal and to report it truthfully. In this section, we derive a demandable-debt contract that achieves this goal.

We make the following assumptions about the population of monitors and the signals:

ASSUMPTIONS: *There are large numbers of potential depositors (Z) and potential monitors (K). The cost of monitoring (I) is small. The probability of any one monitor receiving a bad signal is small. The probability of a bad realization of T is small (although the losses can be large).*

In modeling a bank, each of these assumptions seems natural to us. The assumptions allow us to model the distribution of the number of bad signals as a Poisson distribution. More precise criteria for "small enough" or "large enough" are indicated in the complete appendix (available upon request). Note that as long as I is sufficiently

It can be shown that, for values of S below this range, it will be useful to have two thresholds rather than one. For a range of values of bad signals received, it will be optimal to reduce the promised payment, rather than liquidate the bank. This is analogous to the nuisance contract discussed previously, and as before, it can be precluded by sufficiently high reservation levels of return.

²³Here, reserves are used solely for redistributing payouts between monitors and nonmonitors in an incentive-compatible way. In a richer model, banks would choose between holding reserves and investing more in higher-earning projects.

TABLE 2—PAYOFF TO DEPOSITOR WHO ANNOUNCES g

Project realization	Payoff to depositor announcing g	
	Number of depositors announcing $b < \underline{N}$	Number of depositors announcing $b > \underline{N}$
T_1	$\frac{AT_2 + (Z - Y)S - RN}{Z - N}$	$\frac{MY + (Z - Y)S - RN}{Z - N}$
T_2	$\frac{(Z - Y)S - RN}{Z - N}$	$\frac{MY + (Z - Y)S - RN}{Z - N}$

small, it is always optimal to have all the potential monitors engage in investment.

The contract for all depositors is identical. *Ex post* depositors will pick one of two announcements within the contract. Since there are three information possibilities (observing g , observing b , or not making an investment), there will have to be some pooling in the outcomes. We will build a contract in which it is incentive-compatible for the depositors who have made no investment to pool with those who have observed the good draw.

Each depositor's payoff depends on his announcement and the signal (if any) he observes. We let the symbol $EU(\hat{\sigma}, \sigma)$ denote the expected return for a depositor who observes signal σ and announces signal $\hat{\sigma}$.

Individual depositors are subject to two sorts of constraints: participation constraints (i.e., the contract must give expected returns that are sufficient for depositors to participate) and incentive constraints. From the point of view of the individual depositors, the contract must satisfy the following requirements.

- 1) Always announcing g gives an expected return of S , which exceeds the expected return from always announcing b . This means that depositors with high costs of gathering information will be willing to participate in the contract in the manner specified.
- 2) Announcing the observation truthfully gives a return of $S + I$, which exceeds the return from lying. If conditions in requirement 1 are satisfied as well, then individuals with a cost of I for investing

are willing to make the investment in monitoring and report truthfully.

These constraints for individual depositors can be written as follows:

$$\lambda EU(\hat{g}, g) + (1 - \lambda) EU(\hat{g}, b) \\ = S \geq \lambda EU(\hat{b}, g) + (1 - \lambda) EU(\hat{b}, b)$$

$$\lambda EU(\hat{g}, g) + (1 - \lambda) EU(\hat{b}, b) \\ = S + I \geq \lambda EU(\hat{b}, g) + (1 - \lambda) EU(\hat{g}, b)$$

where λ is the prior probability of signal g .

The scheme we consider has payments of a particularly simple form: any depositor announcing b receives the payment R with certainty. We can call an announcement b a "withdrawal of funds." If more than \underline{N} depositors announce b , the bank is liquidated; otherwise, it is not, and the banker has the option of absconding. In any event, those depositors who do not announce b evenly split the aggregate payment to depositors described in the previous section, less the funds withdrawn. We call this scheme a "standard demandable-debt contract."

Under a standard demandable-debt contract, of course,

$$EU(\hat{b}, b) = EU(\hat{b}, g) = R.$$

However, for depositors who do not withdraw their funds, the payment depends on the number of depositors N who do withdraw, and on whether the banker absconds. Table 2 describes the payments for a depositor who announces g .

For example, if more than \underline{N} depositors withdraw funds, then the bank is liquidated,

and according to the contract, the total payment to depositors P is $MY + (Z - Y)S$; that quantity, less the withdrawn deposits RN is split among the remaining depositors $Z - N$, yielding the quantity in the right-most column of the table. The remaining numbers are calculated in a similar fashion.

Given the probabilities of the realizations of T_i and the probability of each signal contingent on T_i , it is a straightforward matter to calculate $EU(\hat{g}, b)$ and $EU(\hat{g}, g)$. For this scheme, the incentive and participation constraints reduce to the following:²⁴

$$EU(\hat{g}, b) = R - I/(1 - \lambda) \\ S > R.$$

When an aggregate contract of the sort described in the previous section is optimal, it can always be implemented with a demandable-debt scheme, as stated in the following theorem.

THEOREM 4: *Under the distributional assumptions and the conditions of the previous theorem, the optimal outcome can be achieved with a simple demandable-debt contract.*

The role of reserves in our model warrants discussion. By holding reserves, the bank is able to guarantee early payment to a small number of monitors (those who receive bad signals) without forcing the bank to be placed into receivership. Reserves allow the bank to commit to the sequential-service constraint (early withdrawals by those who run the bank), which supports the implementation of the contract between bankers and depositors. More familiar justifications for bank reserve holding include the usefulness of reserves in meeting stochastic demands for conversion into gold

(say, due to foreign-transactions needs of depositors) or the contribution of reserves to an optimally diversified portfolio of bank assets. Our model adds to these transactions and portfolio motivations for holding reserves an "incentive-compatibility" demand for reserves.

C. Transactability and Demandable Debt

Thus far, we have argued that demandable-debt intermediation may arise in order to permit profitable investment opportunities to be realized. In our models, there is no demand for transactability; therefore, assets are valued entirely based on expected return. Historically, however, an important feature of demandable-debt instruments has been their use as a medium of exchange. In this subsection, we briefly consider the implications of our model for the liquidity of demandable debt.

It is important to note from the outset that transactable instruments need not be demandable. Postdated bills of exchange and postdated bank notes were physically transactable instruments that existed in the 19th century in the United States (Davis R. Dewey, 1910). Their primary difference from demandable debt was that they could be redeemed, not on demand, but only on the date of maturity. Since such instruments could be maturity-matched, they would seem to have none of the disadvantages of demandable debt. Nonetheless, demandable debt outcompeted these as a medium of exchange.

In order to explain the relative liquidity of demandable debt, one must explain why the ability to redeem a bank note or deposit on demand makes people more willing to accept it as a means of payment. We argue that, under demandable debt, monitors and nonmonitors alike are better informed of the market value of the debt instrument at all times.²⁵

²⁴The constraints initially have two equalities that must be satisfied. However, given the fact that the total expected payments equal $SZ + KI$, as they do by construction of the demandable-debt contract, one of the equations is redundant: if the informed depositors are each receiving $S + I$, then the uninformed depositors are automatically receiving the remainder, or S per depositor.

²⁵In a different context, Gorton and Pennacchi (1990) also employ this definition of liquidity. They show that debt instruments may be more liquid than equity because debt instruments reduce the potential

The fact that "the bank is open" (that monitors have not called for a liquidation) is revealing to nonmonitors. In the simplest, one-monitor case, the fact that the bank is open is fully revealing, because the signal that the monitor receives takes one of two values. In the multimonitor case, the fact that the bank is open is not fully revealing; it only indicates that fewer than the threshold number of bad signals have been announced. Even this information, however, places a lower bound on the value of the bank's liability.²⁶ If the liquidity of an asset depends on the extent to which information about its value is shared, then one would expect demandable debt to have been more liquid than other contracts with which it competed (see George Akerlof, 1970; Benjamin Klein, 1974). Thus, it may be possible to view the liquidity of bank claims as a by-product of the solution to the agency problem.

While we argue that the transactability of demandable debt enhanced its attractiveness, it is interesting to note that demandable-debt banking predates the transactability of demandable debt.²⁷ Thus, the desirability of demandable-debt contracting does not seem to have depended crucially on the transactability of the instruments.

gains insiders can receive from trading. Their model does not, however, explain the special liquidity of demandable debt.

²⁶Historically, specie prices of bank notes published in bank-note "reporters" confirm the view that nonmonitors faced little price uncertainty for notes of banks that were open. Discounts on antebellum bank notes convertible on demand into specie traded in the home city at par; in distant locations, the discounts for notes mainly reflected the risk due to the time it would take to reach the city of issue. Typically, one could know the value of a bank's notes in New York by knowing the state in which the bank was located. These discounts typically remained small (between $\frac{1}{8}$ percent and 2 percent) and were subject to little variation. Discounts of notes for failed banks were not quoted in bank-note reporters or were subject to extreme variations across banks in the same locale and over time (see Calomiris and Schweikart, 1988).

²⁷For example, Roman banks issued demandable claims which were not transactable (A. W. Ferrin, 1908).

The "liquidity premium" that demandable debt enjoys can be included in our framework by reducing the level of the required return S on demandable debt by the amount of the liquidity premium. In other words, demandable debt would face a lower threshold reservation level to satisfy than the nonliquidating compound contract. This implies an expansion of the parameter values for which demandable debt is preferred over the "nuisance" contract.

IV. Summary

We have argued that historical demandable-debt banking can be understood as the optimal means of incentive-compatible intermediation in an environment of asymmetric information with potential for fraudulent behavior on the part of the banker. Monitoring by some depositors and runs by monitors who receive bad signals ensure sufficiently high payoffs to depositors in states of the world that would otherwise lead to malfeasance by the banker.

Agency problems are inherent in banking. Depositors entrust their endowments to bankers, who decide how to invest them and have essentially unfettered immediate control over the depositors' funds. We capture this agency problem in a simple way by allowing the potential for "absconding" by the banker. The banker has the ability to remove funds from the bank. Absconding is socially wasteful; if the banker steals funds from the bank, he uses a "leaky bucket," so that the amount he actually receives is less than the amount stolen.

If the required return for depositors is sufficiently high, then the banker may find it attractive to abscond, rather than make the promised payment to depositors. Anticipating this, depositors will be unwilling to entrust their funds to the banker, and efficient intermediation will not take place. In other words, the possibility for a banker to abscond may make it difficult for him to attract depositors to his bank.

We introduce a liquidation technology that allows depositors, at a cost, to prevent the banker from absconding and makes it

possible for the banker to attract depositors. We show that, under some circumstances, the optimal arrangement has the depositor choose whether to liquidate the bank, contingent on a costly signal he receives. In good states, it will pay for the banker not to abscond and to pay the depositor as promised; in bad states, absent a liquidation announcement, the banker will abscond rather than pay as promised. Thus, when monitors receive bad signals, they call for liquidation.

If the signal is perfect and costless to the depositor, liquidation will occur only when there are bad loan-investment realizations. If the signal is imperfect and costly, but not prohibitively so, it still makes sense to use the contingent liquidation contract, even though on occasion monitoring depositors may make errors in judging when to "run the bank" and force the bank to liquidate unnecessarily. Banks can fail either because the banker absconds or because the depositor initiates a run on the bank. The purpose of a run is to prevent absconding from taking place.

In the case of multiple depositors, the bank uses reserves to offer guaranteed payments to early withdrawers and to insulate itself from a few bad idiosyncratic signals. At the same time, under circumstances that probably would lead to costly absconding, depositors as a group are likely to order liquidation preemptively. The number of monitors and the threshold at which a bank liquidation is called for will be chosen optimally to minimize total expected costs of liquidation, absconding, and monitoring.

Limitations and Suggested Extensions

Our analysis has several important limitations. First, our goal is to explain the *historical* importance of demandable debt in banking. In today's more regulated environment, where for example, regulations on clearing through the Federal Reserve System have favored demandable-debt instruments and where deposit insurance makes depositor monitoring less important, demandable debt may persist simply as an artifact of regulation.

Second, our framework does not consider the possibility of trade in bank shares. Unlike the historical context in which demandable debt arose, in today's more sophisticated financial markets, shares of financial intermediaries are actively traded. In this richer context, equity trading could conceivably provide a superior disciplinary alternative to demandable debt and contingent liquidation. For example, leveraged buy-outs offer a possible alternative means to prevent managerial misconduct and provide rewards that make monitoring incentive-compatible.

Third, our account is one of individual banks and individual bank liquidations, not of systems of banks or economy-wide bank panics. We are only attempting to model the operation of demandable debt in normal times, when the rules require banks to pay on demand. In historical practice, the provisions of demandable debt, including liquidation, were suspended during crises (see James G. Cannon, 1910; Calomiris and Schweikart, 1988). That is to say, demandable debt was a contingent rule; it required banks to meet the threat of runs in response to idiosyncratic problems, but it allowed banks to escape convertibility on demand in the face of systemic disturbances. Only individual bank difficulties led to placing a bank in receivership. Suspension and interbank relations during panics are important as well, but doing this topic justice requires a larger analysis than the one we have undertaken in this paper (see Calomiris and Kahn, 1989; Gorton, 1989; Calomiris and Gorton, 1990).

Fourth, our model relies on a crude and extremely stylized incentive problem characterized by the "leaky bucket" with which the banker can abscond. This leaky-bucket assumption is useful, because it allows us to model the problem in an extremely simple way, but it raises natural questions as to whether the degree of leakiness necessary to generate the results is at all realistic. After all, if the banker's own stake is less than 1 percent of the value of the assets, then it would be necessary that more than 99 percent of the value of the assets leak from the bucket in good times in order to keep the banker from absconding.

A more reasonable interpretation of our story is as a simplification of a multiperiod account, in which the banker is in fact choosing whether to engage in malfeasance today, when the decision not to engage in malfeasance always leaves the option open for tomorrow. Suppose that the returns to a bank's investments are intertemporally correlated. Then, in a good realization, the banker may be unwilling to engage in malfeasance because it will destroy the prospects for future returns (including the possibility of future malfeasance), even without assuming the bucket implausibly leaky.²⁸ Thus, it is important to investigate multiperiod versions of our model to determine whether a consistent account can be generated with plausible parameter values.

Finally, our model does not include any demand for liquidity. We have intentionally limited the model in order to emphasize the difference between our account and those accounts that depend on liquidity demand. Nonetheless, this limitation means that the model is not adequate to investigate the relation between demandable debt and transactions demand. Although we have briefly and informally considered the links, formal models combining the consumption-flexibility and monitoring accounts of banking are an important goal for future research.

APPENDIX: SKETCHES OF PROOFS

To conserve space, we briefly describe the proofs for each of the four theorems. The complete Appendix is available from the authors on request.

PROOF OF THEOREM 1:

The claim that an optimal contract must conform to one of the four cases listed in the theorem is equivalent to the following claims.

- a) If the promised payment is less than the minimum absconding tax (AT_1), then liq-

uidation is never called for, since absconding is socially wasteful and simple debt repayment is always credibly preferred *ex post* by the banker.

- b) If the optimal contract is a compound contract, then it cannot specify liquidation in all states, since in that case there would be no incentive to invest in signals. If liquidation is going to be called for, it must be that it is only called for under the bad signal.
- c) If the optimal contract involves monitoring and contingent debt claims (the depositor announces one of two values to be repaid), then the amount announced contingent on the bad signal will be lower than the one announced contingent on the good signal, and the lower amount will be less than the minimum absconding tax.

PROOF OF THEOREM 2:

When $S < AT_1$, it is immediate that a simple debt contract is optimal. When the banker chooses between the demandable-debt and nuisance contracts, the banker will always choose the nuisance contract when it is feasible, because it is less socially wasteful than demandable debt. In the nuisance contract, social waste occurs through absconding when a good signal is received but a bad outcome is realized. In the demandable-debt contract, an additional source of waste is the liquidation cost when the bad signal is received. It can be shown that, as the reservation level of the depositor rises, liquidation will eventually be required to increase the depositor's returns beyond what is feasible in the nuisance contract. The use of either form of compound contract requires that the costs of receiving the signal be sufficiently low and the signal's accuracy be sufficiently high to warrant investment in the signal.

PROOF OF THEOREM 3:

The optimal contract is designed to give the depositors their required expected return while minimizing expected social waste from absconding and liquidation. The optimal contract in general involves dividing the possible values of N into three regions. For

²⁸We are grateful to an anonymous referee for suggesting this interpretation.

high values of N , the contract mandates liquidation. For intermediate values of N (a nuisance region), liquidation is not mandated, but aggregate payment is set sufficiently low that absconding never occurs. For low values of N , liquidation is not mandated, and payment is set sufficiently high that absconding takes place in bad states. It can be shown that, as the reservation level of depositors rises, the middle nuisance region disappears, in order to expand the range of higher depositor returns achieved through liquidation or high but uncertain payments.

PROOF OF THEOREM 4:

Given the payoff structure, one can write monitors' and nonmonitors' individual expected returns as functions of the signals received and announced by each, given the probability of other depositors' signals and actions. Tedious but straightforward calculation demonstrates that, for Z and N sufficiently large, the returns so calculated satisfy individual incentive and aggregate feasibility constraints. Finally we show that N sufficiently large can always be found, provided the probability of the good outcome exceeds a certain minimum level.

REFERENCES

- Akerlof, George, "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, August 1970, 89, 488-500.
- Benston, George J. and Kaufman, George G., "Risks and Failures in Banking: Overview, History and Evaluation," Federal Reserve Bank of Chicago, Staff Memorandum SM 86-1, 1986.
- Bernanke, Ben and Gertler, Mark, "Banking and Macroeconomic Equilibrium," in William Barnett and Kenneth J. Singleton, eds., *New Approaches to Monetary Economics* (Proceedings of the Second International Symposium in Economic Theory and Econometrics), Cambridge: Cambridge University Press, 1987, 89-111.
- Bryant, John, "A Model of Reserves, Bank Runs and Deposit Insurance," *Journal of Banking and Finance*, December 1980, 4, 335-44.
- Calomiris, Charles W. and Gorton, Gary, "The Origins of Banking Panics: Models, Facts, and Bank Regulation," working paper, Northwestern University, November 1990.
- _____ and Kahn, Charles M., "A Theoretical Framework for Analyzing Self-Regulation of Banks," working paper, University of Illinois, Urbana, June 1989.
- _____, _____, and Krasa, Stefan, "Optimal Contingent Bank Liquidation Under Moral Hazard," working paper, University of Illinois, Urbana, January 1990.
- _____ and Schweikart, Larry, "Was the South Backward?: North-South Differences in Antebellum Banking During Crisis and Normalcy," working paper, Federal Reserve Bank of Chicago, 1988.
- Cannon, James G., "Clearing Houses," in National Monetary Commission, *Senate Document No. 491, 61st Congress, 2nd Session*, Washington, DC: U.S. Government Printing Office, 1910, pp. 1-335.
- Chari, V. V. and Jagannathan, Ravi, "Banking Panics, Information, and Rational Expectations Equilibrium," *Journal of Finance*, July 1988, 43, 749-63.
- Cone, Kenneth R., "The Regulation of Depository Institutions," Ph.D. Dissertation, Stanford University, 1983.
- Dewey Davis R., "State Banking before the Civil War," in National Monetary Commission, *Senate Document No. 581, 61st Congress, 2nd Session*, Washington, DC: U.S. Government Printing Office, 1910, pp. 1-226.
- Diamond, Douglas W., "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies*, July 1984, 51, 393-414.
- _____ and Dybvig, Philip, "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, June 1983, 91, 401-19.
- Dowd, Kevin, "Option Clauses and the Stability of a Laissez Faire Monetary System," *Journal of Financial Services Research*, December 1988, 1, 319-33.
- Engineer, Merwan, "Bank Runs and the Suspension of Demand Deposit Withdrawals," unpublished manuscript,

- Queen's University, July, 1987.
- Fama, Eugene F.**, "Contract Costs and Financing Decisions," Working Paper 145, Center for Research in Security Prices, University of Chicago, July 1988.
- Ferrin, A. W.**, "The Business Panic of A.D. 33," *Moody's Magazine*, August 1908, 6, 81-2.
- Gendreau, Brian C. and Prince, Scott S.**, "The Private Costs of Bank Failures: Some Historical Evidence," *Federal Reserve Bank of Philadelphia Business Review*, March/April 1986, 3-14.
- Golembe, Carter H. and Warburton, Clark**, *Insurance of Bank Obligations in Six States During the Period 1829-1866*, unpublished manuscript, Federal Deposit Insurance Corporation, 1958.
- Gorton, Gary**, "Self-Regulating Banking Coalitions," working paper, Finance Department, The Wharton School, University of Pennsylvania, June 1989.
- _____ and **Pennacchi, George**, "Transactions Contracts," paper presented at the Garn Institute of Finance Academic Symposium (August 18-20, 1988), Finance Department, The Wharton School, University of Pennsylvania, 1988.
- _____ and _____, "Financial Intermediaries and Liquidity Creation," *Journal of Finance*, March 1990, 45, 49-72.
- Grossman, Sanford J. and Stiglitz, Joseph E.**, "On the Impossibility of Informationally Efficient Markets," *American Economic Review*, June 1980, 70, 393-408.
- Jacklin, Charles J.**, "Demand Deposits, Trading Restrictions, and Risk Sharing," in Edward C. Prescott and Neil Wallace, eds., *Contractual Arrangements for Intertemporal Trade* (Minnesota Studies in Macroeconomics, Vol. 1), Minneapolis: University of Minnesota Press, 1987.
- _____, "Demand Equity and Deposit Insurance," paper presented at the Garn Institute of Finance Academic Symposium (August 18-20, 1988), Stanford University Graduate School of Business, 1988.
- _____ and **Bhattacharya, Sudipto**, "Distinguishing Panics and Information-Based Bank Runs: Welfare and Policy Implications," *Journal of Political Economy*, June 1988, 96, 568-92.
- Klein, Benjamin**, "The Competitive Supply of Money," *Journal of Money, Credit and Banking*, November 1974, 6, 423-53.
- Kyle, Albert S.**, "An Equilibrium Model of Speculation and Hedging," Ph.D. Dissertation, University of Chicago, 1981.
- Nicholas, Henry C.**, "Runs on Banks," *Moody's Magazine*, December 1907, 5, 23-6.
- Smead, E. L.**, "Bank Suspensions in 1927 and During 1921-1927," Unpublished Memorandum, Federal Reserve Board of Governors, 11 April 1928.
- Townsend, Robert M.**, "Optimal Contracts and Competitive Markets with Costly State Verification," *Journal of Economic Theory*, October 1979, 21, 265-93.
- Villamil, A. P.**, "Demand Deposit Contracts, Suspension of Convertibility, and Optimal Financial Intermediation," unpublished manuscript, University of Illinois, Urbana, 1988; *Economic Theory*, forthcoming.
- Comptroller of the Currency**, *Annual Report*, Washington, DC: U.S. Government Printing Office, 1920.
- _____, "An Evaluation of the Factors Contributing to the Failure of National Banks," unpublished memorandum, January 1988.

A Model of Homogeneous Input Demand Under Price Uncertainty

By FRANK A. WOLAK AND CHARLES D. KOLSTAD*

This paper examines the empirical validity of a model of homogeneous input demand under price uncertainty in which firms trade off expected input cost against its variability (risk) in selecting the optimal input supplier mix. Using recent work in time-series econometrics, this model is applied to the Japanese steam-coal import market, where five suppliers compete: China, the Soviet Union, South Africa, the United States, and Australia. (JEL L10, L72)

The purpose of this paper is to derive and examine the empirical validity of a model of homogeneous input demand under price uncertainty. The motivation for this investigation is the common observation that firms simultaneously purchase a homogeneous factor of production from a variety of suppliers each charging a different price. Moreover, there are many instances when the price from one supplier is consistently above that of all other suppliers for an extended period of time yet firms continue to purchase from this supplier. This observation appears to violate the criterion of expected cost minimization for input choice.¹ An at-

tempt to explain these anomalies suggests that firms trade off the level of expected input cost against its variability in deciding how to allocate total input demand across available suppliers. By purchasing inputs from a variety of suppliers, the firm is diversifying away some of the price risk associated with satisfying demand from the single least-expected-cost supplier.²

Although the marginal rate of substitution (MRS) between risk and cost is not directly observable, we develop a methodology for empirically estimating this magnitude from a time-series of input purchases. This MRS is an estimate of the firm's risk preferences at the expected cost-risk pair selected. If we assume that this MRS between risk and cost is constant across all expected cost-risk pairs, then an input-price risk premium can be calculated. Subject to this assumption, the input-price risk premium is the percentage above the current

*Department of Economics, Stanford University, Stanford, CA 94305, and Department of Economics and Institute for Environmental Studies, University of Illinois, Urbana, IL 61801, respectively. We thank seminar participants at Stanford University, the University of California-Berkeley, the University of Texas, the University of Washington, Purdue University, and the Norwegian School of Economics for comments on earlier drafts. Tom MaCurdy, Randy Mariger, Paul Newbold, Roger Noll, and Agnar Sandmo deserve special mention for their helpful comments. Vivian Hamilton expertly prepared the figures. We especially thank an anonymous referee for thoughtful comments and suggestions on the previous version of the paper. His many contributions are too numerous to mention individually. The final version of this paper was prepared while Wolak was a National Fellow of the Hoover Institution.

¹For the sake of simplicity, assume that the price series are independent and identically distributed draws from a multivariate distribution. The null hypothesis of equal means for the prices becomes less likely the

greater the number of observations that one price series remains above the others. Clearly, if firms are minimizing expected cost, they would purchase all of this input from the least-expected-price supplier. Hence, in this simple case, the nonzero market share of the consistently high-priced supplier is, with high probability, a violation of the expected-cost-minimization criterion of input choice.

²Previous authors (Agnar Sandmo, 1971; Raveendra N. Batra and Aman Ullah, 1974; Roger D. Blair, 1974) have theoretically examined the comparative statistics of firm behavior under input and output price uncertainty.

TABLE 1—SUMMARY STATISTICS FOR PRICES AND QUANTITY SHARES
(SAMPLE PERIOD: MAY 1983–MAY 1987)

Country	Mean price. ($\frac{10^3 \text{ yen}}{\text{metric ton}}$)	Standard deviation of price	Mean quantity share	Standard deviation of quantity share
China	10.49	2.60	0.087	0.025
Soviet Union	8.95	1.67	0.036	0.015
United States	14.00	3.07	0.119	0.052
South Africa	10.57	2.39	0.213	0.054
Australia	10.54	2.45	0.547	0.087

Data Source: Japan Export and Imports: Commodity by Country.

expected market price a firm would pay for riskless input supply. If the firm's preferences imply a declining MRS between risk and expected cost (the MRS depends on the level of these two magnitudes), then the risk premium we compute is only an upper bound on the percentage above the current expected price the firm would be willing to pay for riskless input supply. Our risk-diversification model of input demand also provides a framework for quantifying the relative risk characteristics of input prices similar to the framework for assessing the relative risk of securities in the capital-asset-pricing model (CAPM). This framework will be discussed later in the paper.

We have chosen the Japanese steam-coal import market for an empirical implementation of the risk-diversification model. This coal is primarily used in Japanese cement-manufacturing and electricity-generation facilities. Although this coal is supplied to a variety of consumers in Japan, the Ministry of International Trade and Industry (MITI) is the centralized decision-maker which coordinates all international steam-coal transactions and hence is analogous to the firm in our model of input demand. We estimate the risk-diversification model and examine its validity as an explanation for the observed patterns of Japanese steam-coal imports.

The specific puzzle we address is: why do the Japanese not buy the least-expected-cost coal? Three observations about the time-series properties of the vector of yen prices and quantities of steam coal imported from

the five suppliers—China, the Soviet Union, the United States, South Africa, and Australia—provide evidence against the expected-cost-minimization model of input choice. Table 1 gives the mean price in thousands of yen per metric ton and the mean quantity share, as well as the standard errors for both of these quantities, for these five countries over our sample period.

The first observation is that the price of United States coal is above that of all other suppliers throughout the entire sample period, yet the United States supplies an average of 11.9 percent of all steam coal imported to Japan during this period. The second observation is that the price of steam coal from the Soviet Union is consistently below the price of all other suppliers throughout the sample although it consistently has the smallest share of the Japanese steam-coal import market. These two observations are confirmed by the sample means of the prices given in Table 1. The final puzzle is that South Africa and Australia have approximately the same mean price over the sample, although for all observations over this same period the share of Japanese steam-coal imports from Australia is consistently more than double that from South Africa. We find that the risk-diversification model and the apparent risk characteristics that it implies for each supplier provide an economically plausible explanation of the operation of the Japanese steam-coal import market.

The remainder of the paper proceeds as follows. The next section introduces nota-

tion and then derives the risk-diversification model of input demand. Section II discusses the econometric framework underlying the estimation of this model. This section treats the specification of a stochastic process describing the behavior of the vector of input prices over time and also describes the form and sources of other uncertainty in the model. Section III provides a brief overview of the Japanese steam-coal import market, to match up the theoretical model of Section I with the actual workings of this market. Section IV describes the application of this framework to the Japanese steam-coal import market. It presents several formal and informal tests of our structural model embodying the risk-diversification hypothesis. In Section V, we present the general implications of modeling input demand under uncertainty within this risk-diversification framework. For example, we are able to calculate the risk premium described earlier and a measure of market-specific risk associated with each of the supply-price processes. We can also derive a relationship between these measures of market-specific risk and the optimal expected supply price for each supplier. We then examine the validity of these implications of our structural model within the context of the Japanese steam-coal market. The paper closes with a short discussion of the policy implications of the empirical results and suggestions for future applications of this framework.

I. A Risk-Diversification Model of Firm Input Demand

Consider a firm using a set of inputs to produce one or more outputs. All inputs to production but one are termed "nonrisky" in that their price is nonstochastic. Output prices are also nonstochastic. The price of one of the inputs (the "risky" input) is uncertain. Supplies of that input must be contracted for *ex ante* before the price uncertainty is resolved. If the firm is risk-averse, it may increase its utility by substituting away from the risky input or by utilizing a variety of suppliers in an effort to reduce risk through diversification.

We begin by defining notation:

- p_{it} : price of risky input from supplier i in period t ($i = 1, \dots, n$);
- q_{it} : quantity of risky input demanded from supplier i in period t ($i = 1, \dots, n$);
- \mathbf{p}_t : n -dimensional vector of risky-input prices in period t ;
- \mathbf{q}_t : n -dimensional vector of risky-input quantities demanded in period t ;
- \mathbf{r}_t : vector of prices of nonrisky inputs in period t ;
- \mathbf{s}_t : vector of quantities of nonrisky inputs in period t ;
- $\boldsymbol{\pi}_t$: vector of deterministic output prices in period t ;
- \mathbf{y}_t : vector of output quantities in period t ;
- I_t : information set available to firm at time t , containing \mathbf{p}_s ($s \leq t-1$);
- μ_t : $E(\mathbf{p}_t | I_t)$, conditional expectation of \mathbf{p}_t ;
- Σ_t : $E\{(\mathbf{p}_t - \mu_t)(\mathbf{p}_t - \mu_t)' | I_t\}$, conditional variance of \mathbf{p}_t ;
- $\mathbf{1}$: n -dimensional vector of 1's;
- Q_t : $\mathbf{1}'\mathbf{q}_t$, total demand for risky input in period t ;
- \mathbf{w}_t : \mathbf{q}_t / Q_t , n -dimensional vector of risky-input quantity shares.

The firm is governed by the implicit production relation $f(\mathbf{y}_t, \mathbf{s}_t, Q_t) = 0$. Rather than maximize profits, because it is risk-averse, in each period the firm maximizes the expected utility of profits given the vectors of nonstochastic input and output prices and the information set I_t . We make the simplifying assumption that the firm's expected utility can be written as a function of the conditional expectation of profits, $E(\Pi_t | I_t)$, and the conditional variance of profits, $V(\Pi_t | I_t)$, where

$$\Pi_t = \boldsymbol{\pi}_t' \mathbf{y}_t - \mathbf{p}_t' \mathbf{q}_t - \mathbf{r}_t' \mathbf{s}_t$$

is the firm's profit in period t . This assumption about firm preferences is similar to that made for investor preferences in the CAPM. As in the CAPM, this assumption is equivalent to either the firm having a utility function that is quadratic in profits or the random input prices \mathbf{p}_t having a multivariate Gaussian distribution. Thus, the firm's prob-

lem is, at every time period,

$$(1) \max_{\mathbf{q}_t, \mathbf{s}_t, \mathbf{y}_t} U[E(\pi'_t \mathbf{y}_t - \mathbf{p}'_t \mathbf{q}_t - \mathbf{r}'_t \mathbf{s}_t | I_t), \\ V(\pi'_t \mathbf{y}_t - \mathbf{p}'_t \mathbf{q}_t - \mathbf{r}'_t \mathbf{s}_t | I_t)] \\ \equiv U[\pi'_t \mathbf{y}_t - \mathbf{r}'_t \mathbf{s}_t - E(\mathbf{p}'_t \mathbf{q}_t | I_t), \\ V(\mathbf{p}'_t \mathbf{q}_t | I_t)]$$

subject to

$$f(\mathbf{y}_t, \mathbf{s}_t, Q_t) = 0, \quad \mathbf{v}'_t \mathbf{q}_t = Q_t, \quad \mathbf{q}_t, \mathbf{s}_t, \mathbf{y}_t \geq 0.$$

This optimization problem is equivalent to the two-stage process whereby first an optimal portfolio of suppliers is chosen to yield a given Q_t . Then, in the second stage, the proper balance is struck among outputs (\mathbf{y}_t), nonrisky inputs (\mathbf{s}_t), and the total amount of the risky input (Q_t). The portfolio of \mathbf{q}_t for a given Q_t , and F (described below) is the solution to

$$(2) \max_{\mathbf{q}_t} U[F - E(\mathbf{p}'_t \mathbf{q}_t | I_t), V(\mathbf{p}'_t \mathbf{q}_t | I_t)] \\ \text{subject to } \mathbf{v}'_t \mathbf{q}_t = Q_t, \quad \mathbf{q}_t \geq 0.$$

Substituting this vector of optimal supplier quantities back into the objective function yields the optimal-value function $U^*(F, Q_t | I_t)$, where F is net revenue from nonrisky inputs and outputs. Thus, U^* defines the highest level of utility obtainable for a given F and Q_t ; the optimal \mathbf{q}_t^* (which is a function of F and Q_t) has been substituted in for \mathbf{q}_t . The second-stage optimization problem uses this optimal-value function to determine the utility-maximizing total quantity of the risky input (Q_t), nonrisky inputs (\mathbf{s}_t), and outputs (\mathbf{y}_t) as follows:

$$(3) \max_{\mathbf{y}_t, \mathbf{s}_t, Q_t} U^*(\pi'_t \mathbf{y}_t - \mathbf{r}'_t \mathbf{s}_t, Q_t | I_t)$$

subject to

$$f(\mathbf{y}_t, \mathbf{s}_t, Q_t) = 0, \quad Q_t, \mathbf{s}_t, \mathbf{y}_t \geq 0.$$

Solving (3) with U^* defined by (2) is equivalent to solving (1).

Because we are only interested in the choice of the portfolio of suppliers of the risky input, we will focus on (2). Ignoring the possible negativity of any elements of \mathbf{q}_t , the Lagrangian for (2) is

$$(4) \quad L = U(F - \mu'_t \mathbf{q}_t, \mathbf{q}'_t \Sigma_t \mathbf{q}_t) \\ + \eta(Q_t - \mathbf{v}'_t \mathbf{q}_t)$$

where η is the Lagrange multiplier on the constraint that the sum of purchases from all of the suppliers equals Q_t . The first-order conditions from (4) are

$$(5) \quad \frac{\partial L}{\partial \mathbf{q}_t} = -U_1 \mu'_t + 2U_2 \mathbf{q}'_t \Sigma_t - \eta \mathbf{v}' = 0$$

where U_i is the derivative of U with respect to its i th argument. Equation (5) can be solved for the scalar η using the constraint $\mathbf{v}'_t \mathbf{q}_t = Q_t$:

$$(6) \quad \eta = \frac{2U_2 Q_t - U_1 \mathbf{v}' \Sigma_t^{-1} \mu_t}{\mathbf{v}' \Sigma_t^{-1} \mathbf{v}}$$

Substituting (6) back into (5) and rearranging gives the following expression for the optimal vector of risky input shares:

$$(7) \quad \mathbf{w}_t^0 = \left[\frac{\lambda_t Q_t + \mathbf{v}' \Sigma_t^{-1} \mu_t}{\mathbf{v}' \Sigma_t^{-1} \mathbf{v}} (\Sigma_t^{-1} \mathbf{v}) - \Sigma_t^{-1} \mu_t \right] \frac{1}{\lambda_t Q_t}$$

where $\lambda_t = -2U_2/U_1$. Dividing λ_t by 2 gives the producer's marginal rate of substitution between expected costs and risk. It is, of course, a function of I_t , Q_t , and F . However, Q_t and F , and thus λ_t , are the result of solving (3). Rather than solve (3) explicitly, we make an assumption about the functional form of λ_t . Several specifications for λ_t are possible. The first is simply $\lambda_t = \lambda$ for all t . Another, which is the specification we adopt, is that $\lambda_t = \lambda/Q_t$ (i.e., $\lambda_t Q_t$ is a constant). This specification for λ_t has the attractive feature that it makes the optimal input share (7) invariant to Q_t . Thus, for fixed μ_t and Σ_t , if there is a secular rise in

the level of productive activity in the firm, supplier shares remain constant. This expression for λ_t simplifies (7) to

(8)

$$\mathbf{w}_t^o = \left[\frac{\lambda + \mathbf{v}' \Sigma_t^{-1} \boldsymbol{\mu}_t}{(\mathbf{v}' \Sigma_t^{-1} \mathbf{v})} (\Sigma_t^{-1} \mathbf{v}) - \Sigma_t^{-1} \boldsymbol{\mu}_t \right] \frac{1}{\lambda}.$$

For notational ease in what follows we write (8) as

$$(9) \quad \mathbf{w}_t^o = \mathbf{S}_t(\boldsymbol{\mu}_t, \Sigma_t, \lambda).$$

Defining $\phi = 1/\lambda$, we can rewrite (8) as

(10)

$$\mathbf{w}_t^o = \left[\frac{1 + \phi(\mathbf{v}' \Sigma_t^{-1} \boldsymbol{\mu}_t)}{(\mathbf{v}' \Sigma_t^{-1} \mathbf{v})} (\Sigma_t^{-1} \mathbf{v}) - \phi(\Sigma_t^{-1} \boldsymbol{\mu}_t) \right].$$

Equation (10) gives the optimal-supplier-share vector as a function of ϕ , the relative weight attached to expected cost in the firm's optimal-input-choice problem, whereas (8) gives the same optimal supplier share as a function of λ , the relative weight attached to the conditional variance of input cost. Equation (10) will be used later when we examine the validity of our structural model of the risk-diversification hypothesis.

Given values for $\boldsymbol{\mu}_t$ and Σ_t , and knowing its value of λ , or equivalently ϕ , the firm can compute the optimal period- t input-supplier mix from equation (8). We assume that the firm knows or behaves as if it knows the parameters of the stochastic process determining the time path of the vector of input prices so that it can compute $\boldsymbol{\mu}_t$ and Σ_t for all t . Unfortunately, in order for us to implement this model and determine its empirical validity, we must estimate the parameters of this stochastic process. Therefore, we now turn to the econometrics of the risk-diversification model of input demand.

II. The Econometrics of the Risk-Diversification Model of Input Demand

In this section, we present our methodology for implementing the risk-diversification

model of input demand. There are two independent sources of uncertainty in this model. The first, what we call estimation error, arises from the estimation of $\boldsymbol{\mu}_t$ and Σ_t , the conditional mean and covariance matrix of the vector-valued price process. The second, what we refer to as optimization error, is included to account for any unobservable time-specific random shocks which may cause the first-order conditions (5) not to hold exactly each period.

This optimization error has the implication that we require the first-order conditions to hold only in expectation. Operationally, this means that (9) becomes

$$(11) \quad \mathbf{w}_t = \mathbf{S}_t(\boldsymbol{\mu}_t, \Sigma_t, \lambda) + \boldsymbol{\varepsilon}_t \equiv \mathbf{w}_t^o + \boldsymbol{\varepsilon}_t$$

where $\boldsymbol{\varepsilon}_t \in R^n$ is $\mathcal{N}(0, \Omega)$. Thus, the observed input-share vector (\mathbf{w}_t) equals the optimal input-share vector (\mathbf{w}_t^o) plus white noise. The restriction that $\mathbf{v}' \mathbf{w}_t = 1$ implies that $\mathbf{v}' \boldsymbol{\varepsilon}_t = 0$ and $\mathbf{v}' \Omega \mathbf{v} = 0$. We introduce this optimization error as a way to take into account the fact that there may be unobservable (to the econometrician) variables that affect the supplier shares actually selected which are not included in our model of input demand. Hence, despite requiring the total input supply to be deterministic, the model does allow random variation in quantities across suppliers from the utility-maximizing levels determined from our risk-diversification model.

Because our price series appear to be stationary in first-differences yet their levels roughly move together (there exists a stationary linear combination of the prices), our methodology for modeling the estimation error in $\boldsymbol{\mu}_t$ and Σ_t involves fitting an error-correction model for each price series:

$$(12) \quad \Delta p_{it} = c_i + \gamma_i z_{it-1} + \beta_i \Delta p_{it-1} + \xi_{it} \\ (i = 1, \dots, n; t = 1, \dots, T)$$

where $\Delta p_{it} = p_{it} - p_{it-1}$. To take into account the fact that, on average, the levels of these prices move together we include z_{it} , which is an estimate of the stationary linear combination of all of the prices. Following

Robert F. Engle and Clive W. J. Granger (1987) and James H. Stock (1987), we compute z_{it} as the residual from the regression of p_{it} on all other prices and a constant. Hence, z_{it} has a sample mean of zero. As shown in Stock (1987), because the parameters of the cointegrating regression converge to their true values at rate T , rather than the usual \sqrt{T} , the z_{it} may be effectively treated as the observed z_t in the estimation of the parameters of (12) and the computation of their \sqrt{T} -asymptotic distribution. We assume that $\xi_t = (\xi_{1t}, \xi_{2t}, \dots, \xi_{nt})'$ is distributed as a $\mathcal{N}(\mathbf{0}, \Sigma)$ random vector. This distributional assumption for ξ_t and the model (12) for p_{it} ($i = 1, \dots, n$) implies that the conditional variance of Σ_t equals a constant Σ for all t .

Besides embodying the cointegration property of \mathbf{p}_t , this model for each price series is consistent with the following logic. The constant term c_i takes into account the possibility that Δp_{it} may have a nonzero mean. By including this constant term, we are effectively allowing p_{it} to have a nonzero mean in its stochastic trend. The term in z_{it-1} , the error correction term, takes into account the fact that the amount z_{it-1} differs from its steady-state value will affect period t 's price change for this supplier. We would expect γ_i to be negative because, if z_{it-1} is positive (recall how z_{it} is estimated), this period's Δp_{it} should be lower than its mean to reflect a correction toward the steady state. The third term in Δp_{it-1} represents the impact of last period's price change on this period's price change.

The final model estimated for each price series should be such that the null hypothesis that $\xi_{it} \equiv E(p_{it} | I_t) - p_{it}$ is white noise cannot be rejected. Additional terms in Δp_{js} ($j = 1, \dots, n$; $s \leq t-1$) should be added to the model until this is the case. Clearly, a shortcoming of our approach is that there may be other variables besides lagged values of p_t that help to estimate μ_t . Nevertheless, in this paper we assume that I_t contains only lagged values of \mathbf{p}_t .

Let Γ_i denote all of the coefficients entering into (12) for Δp_{it} . Let $\mu_{it}(\Gamma_i, I_t)$ denote the conditional mean function for the i th price process. In this shorthand notation we

can rewrite (12) as

$$(13) \quad p_{it} = \mu_{it}(\Gamma_i, I_t) + \xi_{it} \quad (i = 1, \dots, n).$$

If we stack all of the Γ_i into a single vector Γ then we can write (13) in vector notation as:

$$(14) \quad \mathbf{p}_t = \mu_t(\Gamma, I_t) + \xi_t.$$

Once we fit a univariate model to each price series such that the null hypothesis that each ξ_{it} series is white noise cannot be rejected, we can construct a consistent estimate of Σ as follows:

$$\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^T \tilde{\xi}_t \tilde{\xi}_t'$$

where $\tilde{\xi}_t = (\tilde{\xi}_{1t}, \tilde{\xi}_{2t}, \dots, \tilde{\xi}_{nt})'$ and $\tilde{\xi}_{it}$ is the ordinary least-squares (OLS) estimate of ξ_{it} . This completes the first step of our two-step procedure to obtain consistent estimates of Σ and Γ . In summary, the equation-by-equation OLS estimates of the Γ_i in (13) yield a consistent estimate of Γ , and because $\hat{\Sigma}$ is based on this estimate of Γ , it is also consistent. The next step of this procedure conditions on these estimates of Γ and Σ and computes [from (11)] estimates of λ and Ω by maximum likelihood (ML). This two-step process yields \sqrt{T} -consistent estimates of Γ , Σ , λ , and Ω which can be used as starting values in a full-model ML estimation procedure.

Combining the model determining \mathbf{w}_t in (11) with that determining \mathbf{p}_t in (13) yields the following nonlinear ML model:

$$(15) \quad \begin{bmatrix} \mathbf{p}_t \\ \mathbf{w}_t \end{bmatrix} = \begin{bmatrix} \mu_t(\Gamma, I_t) \\ S_t(\mu_t(\Gamma, I_t), \Sigma, \lambda) \end{bmatrix} + \begin{bmatrix} \xi_t \\ \epsilon_t \end{bmatrix}$$

where $E(\xi_t \epsilon_t') = 0$, because the estimation error is assumed to be independent of the optimization error.

The log-likelihood function is

$$\begin{aligned}
 (16) \quad \ln L = & -\frac{T(2n-1)}{2} \ln 2\pi - \frac{T}{2} \ln[\det(\Sigma)] \\
 & - \frac{1}{2} \sum_{t=1}^T \{[\mathbf{p}_t - \boldsymbol{\mu}_t(\boldsymbol{\Gamma}, \mathbf{I}_t)]' \\
 & \quad \times \boldsymbol{\Sigma}^{-1}[\mathbf{p}_t - \boldsymbol{\mu}_t(\boldsymbol{\Gamma}, \mathbf{I}_t)]\} \\
 & - \frac{T}{2} \ln[\det(\boldsymbol{\Omega})] \\
 & - \frac{1}{2} \sum_{t=1}^T \{[\mathbf{w}_t - \mathbf{S}_t(\boldsymbol{\mu}_t(\boldsymbol{\Gamma}, \mathbf{I}_t), \boldsymbol{\Sigma}, \boldsymbol{\lambda})]' \\
 & \quad \times \boldsymbol{\Omega}^{-1}[\mathbf{w}_t - \mathbf{S}_t(\boldsymbol{\mu}_t(\boldsymbol{\Gamma}, \mathbf{I}_t), \boldsymbol{\Sigma}, \boldsymbol{\lambda})]\}.
 \end{aligned}$$

Given the two-step \sqrt{T} -consistent estimates of $\boldsymbol{\Gamma}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\lambda}$ described above, by the logic of theorem 6.3.1 of Erich L. Lehmann (1983 p. 422), asymptotically efficient estimates of these parameters can be obtained by one iteration of a method-of-scoring type of algorithm. Alternatively, starting from these consistent estimates and running this iterative procedure to convergence also yields asymptotically efficient estimates of these parameters.³

³We use the procedure suggested by Ernst R. Berndt et al. (1974) to compute the iterative maximum-likelihood estimates of these parameters. To simplify the computational complexity of the problem, we estimate $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ in terms of the parameters of the Cholesky decomposition of their inverses. Recall that $\boldsymbol{\Sigma}^{-1}$ can be written as \mathbf{LDL}' , where \mathbf{L} is a lower triangular matrix with 1's along the diagonal and \mathbf{D} is a diagonal matrix. The determinant of $\boldsymbol{\Sigma}^{-1}$ is the product of the diagonal elements of \mathbf{D} . This decomposition simplifies the terms in the log-likelihood function containing the determinant of $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ to a product of four diagonal elements in the former case and the product of five diagonal elements in the latter case. By the invariance property of maximum-likelihood estimation, the maximum-likelihood estimates of $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ are equal to the inverse of the maximum-likelihood estimates of the Cholesky decomposition of the parameters of their respective inverse matrices. Consistent estimates of the standard errors can be obtained from the sample average of the matrix of outer products of the gradients of the log-likelihood function evaluated at the maximum-likelihood estimate of the parameter vector as described in Berndt et al. (1974) or as -1 times the matrix of second partial derivatives of the log-likelihood function evaluated at the same value of the parameter vector.

Given this framework for specifying and estimating our model of input choice under price uncertainty, we are now ready to apply it to the Japanese steam-coal import market. Before proceeding to the application, we first describe the history and operation of this market.

III. The Japanese Steam-Coal Import Market

Almost immediately after the 1973–1974 Arab Oil Embargo and the subsequent substantial increase in the world price of oil, the Japanese embarked on a plan, coordinated between business and government, for a stable domestic energy supply (Yuan-li Wu, 1977). Foremost among the methods Japan used to achieve this goal was to diversify both the suppliers and sources of energy. Prior to this event, Japan had an oil-based energy economy and obtained most of its oil from the Middle East and the United States. Following this embargo, Japan expanded its sources of oil to China and the Soviet Union and began to consider coal as a major source of energy.

At this time, Japan was importing coal primarily from the United States for use as coking coal in the production of steel. By the beginning of 1977, the Soviet Union, China, South Africa, and Australia had become consistent participants in this market, but the United States was still the major source of Japanese coal imports. By this time, Japan was also importing steam coal to be burned in coal-fired electricity-generation facilities. In response to the oil price rises, Japan quickly converted most of its cement-manufacturing plants from oil-fired to steam-coal-fired (Ercan Tukenmez and Nancy Tuck, 1984). During the next five years, the United States' share of the steam-coal market steadily declined, and the shares of South Africa and Australia increased considerably. The average volume of monthly steam-coal imports (as classified by the Japan Tariff Association) rose from approximately 300,000 metric tons per month in early 1977 to 1.5 million metric tons per month in early 1984 and eventually to close to 2.7 million metric tons per month in mid-1987.

This steam coal is imported through negotiations with Japanese trading companies in conjunction with MITI for delivery to the steam-coal-using facilities. Prices for coal are negotiated in terms of the currency of the country of origin of the coal, although sometimes in dollars. Hence, some of the price risk borne by Japanese consumers is due to foreign-exchange-rate risk. An additional source of price uncertainty to Japan arises from what are called demurrage costs. These costs are incurred when a ship picking up or delivering coal is unable to load or unload its cargo immediately upon arrival at port. These queuing costs at port are usually directly added to the delivered price of coal. In periods when loading or unloading facilities are operating at capacity, these charges can amount to a significant portion (approximately 10 percent) of the delivered price of coal (United States Department of Energy, 1981 p. 7).

Coal is purchased using three mechanisms: joint venture between buyer and supplier, long-term contract, and short-term supply agreement (this category includes spot-market purchases). In contrast to metallurgical coal, which is a highly specialized input to the production of steel and as a consequence is, for the most part, delivered on long-term contracts, the relatively simple uses of steam coal and the increasing flexibility of boilers to burn different types of coal make short-term supply agreements (less than one year) and spot-market purchases viable. The Japanese make substantial spot-market purchases from South Africa, the United States, and Australia, while spot purchases play a lesser role in the imports from China and the Soviet Union. As stated in a recent United States Department of Energy report, "At present, South Africa steam-coal exports to Asia are largely spot sales" (Tukenmez and Tuck, 1984 p. xxi).

Most long-term contracts and joint ventures allow for some flexibility in the prices charged for coal delivered on a given contract depending on current market conditions at the time of delivery, so that some of the price risk is due to the conditions in the spot market at the delivery date. Also, in

the case of long-term contracts and joint ventures, the Japanese trading companies often renegotiate the prices charged under these agreements if current market conditions favor their doing so. For example, when there is a downturn in the world coal market many of these contracts are renegotiated. This potential for renegotiation of long-term supply agreements based on current market conditions is another source of price uncertainty.

There is an abundance of anecdotal evidence for the validity of the risk-diversification model of input choice for the Japanese steam-coal import market. Various editions of the *MITI Handbook* published by Japan Trade and Industry Publicity, state that the two major policy goals for MITI in the area of energy and natural resources are: 1) a stable supply of energy resources and 2) stable prices of energy resources. One of the stated goals of the Coal Mining Department of MITI is "to smooth the importation of coal" (*MITI Handbook* 1979/1980 p. 82). Japan's desire for a stable, secure energy supply is well documented in Wu (1977), a study of Japan's response to the Arab Oil Embargo of 1973/1974. In addition, a U.S. Department of Energy study of coal trade in the Asian market states, "...in seeking diversification and security Japan seems willing to pay a premium to access stable coal supplies from the more expensive exporters, such as the United States..." (Tukenmez and Tuck, 1984 p. 3). This casual evidence coupled with the three puzzles concerning the time-series properties of the prices and quantities of imports of steam coal to Japan stated in the Introduction makes for a challenging application of our risk-diversification model of input choice that is also of substantial policy interest.

IV. Application to the Japanese Steam-Coal Import Market

Time-series of prices and quantities of steam coal⁴ imported into Japan from

⁴Steam coal is classified by the Japan Tariff Association as high- and low-ash coal other than coking coal.

China, the Soviet Union, the United States, South Africa, and Australia are available on a monthly basis from *Japan Exports and Imports: Commodity by Country*, compiled by the Japan Tariff Association. All prices are in units of thousands of yen per metric ton. The quantity units are metric tons. The Appendix describes the construction of these magnitudes from the raw data. Note that the input-choice problem is invariant to the absolute price level. The normalization of prices will only affect the magnitude of λ . To make shares and prices of approximately the same magnitude in the estimation procedure, prices were normalized so that the sum of the sample means of all of the prices of coal is equal to 1.

The sample period from March 1983 to May 1987 was selected because the structure of the Japanese steam-coal import market seems stable over this period. Confirmation of this point is that, despite a growing total quantity of steam coal imported, the share of the market served by each supplier shows no statistically significant serial correlation or trend over this period. This empirical observation provides further support for our selection of a form for λ , that makes the optimal supplier shares independent of Q_t because, as mentioned in Section III, Q_t nearly doubled over our sample period.

The first step of the estimation procedure is to test for cointegration among the five price processes over the sample. As discussed in Engle and Granger (1987), the presence of cointegration is necessary for

the validity of the error-correction model of the price processes given in (12). To confirm that each of the univariate price processes is integrated of order one, we performed David A. Dickey and Wayne A. Fuller's (1979) unit-root tests on the levels and first differences of each series. The models run for each test are

$$(17) \quad \Delta p_{it} = \alpha + \beta_1 p_{it-1} + \beta_2 \Delta p_{it-1} + e_{it}$$

for the test for a unit root in the levels and

$$(18) \quad \Delta^2 p_{it} = \alpha + \beta_1 \Delta p_{it-1} + \beta_2 \Delta^2 p_{it-1} + d_{it}$$

for the test for a unit root in the first differences. In both cases, the null hypothesis is that $\beta_1 = 0$, or more precisely, the backshift operator polynomial of the AR portion of the ARIMA representation of x_t (x_t represents either the raw or first-differenced price series) has the following factorization: $\phi(B) = (1 - B)\phi^*(B)$, where all of the roots of $\phi^*(z) = 0$ are greater than 1 in modulus. The results of these tests are given in Table 2. For all of the tests in terms of the levels of the price series, there is little evidence against the null hypothesis of a unit root, indicating that nonstationarity of the price series in levels cannot be rejected. In contrast, the null hypothesis of a unit root in the first-differenced series is decisively rejected for all of the series at the 0.01 level of significance, providing strong evidence for the stationarity of the first-differenced series. The critical value for the test is from table 8.5.2 of Fuller (1976 p. 373). For the present case, an assumption implicit in the Dickey-Fuller test—that the true value of α is 0 in (17) and (18)—may not be valid given the substantial decline in prices over the sample period. For this reason, we also report Kenneth D. West's (1988) corrected t statistic on β_1 in (17) and (18). This statistic is asymptotically normal under the assumption that α in these two equations is nonzero. Computing West's t statistic amounts to correcting the usual OLS t statistic for the fact that the OLS estimate of the variance of the error term

Although, strictly speaking, steam coal differs across countries, it is primarily, if not exclusively, valued for its heat content. Consequently, only coal with the highest heat content is exported. Although the heat content of each shipment of coal to Japan during the sample period was not available, the heat content of coal for a representative sample of coal contracts from each of the supplier countries considered in this paper was available (TEX Report, 1986). For this representative sample, the mean heat content per ton of coal delivered was not significantly different across the supplier countries considered here. This provides support for our treatment of steam coal from various countries as a homogeneous product.

TABLE 2—DICKEY-FULLER (DF) AND WEST (W) TESTS FOR UNIT ROOTS

Country	p_{it} (DF)	Δp_{it} (DF)	p_{it} (W)	Δp_{it} (W)
China	-0.8871	-6.0236	-0.9184	-6.1560
Soviet Union	-0.5911	-6.0028	-0.6136	-6.1802
United States	-0.8124	-7.2048	-0.8452	-7.4585
South Africa	-0.0488	-5.4708	-0.0504	-5.5622
Australia	0.4913	-5.6874	0.5051	-5.7431

Notes: Critical value (0.01 level) for Dickey-Fuller test = -3.58; critical value (0.01 level) for West test = -2.33.

e_{it} in (17) or d_{it} in (18), is inconsistent if these errors are autocorrelated. West suggests a consistent estimate of this variance based on the sample autocorrelation function of the OLS residuals. To compute West's t statistic, we must first choose m , the number of sample autocorrelations to include in the estimate of the variance of the error in (17) or (18). We selected $m = 5$ because beyond this value of m the value of \hat{s} (in West's notation) did not appreciably change. These statistics are reported in the second column of Table 2. The one-sided critical value for these statistics is obtained from the standard normal distribution; the results of West's tests confirm the results of the Dickey-Fuller tests. This battery of tests is in line with the first requirement for the price processes to be cointegrated. The results suggest that each of the univariate price processes is integrated of order one.

The second requirement of cointegration is that some linear combination of the prices is stationary. To test this hypothesis, we utilize the augmented Dickey-Fuller (ADF) test recommended by Engle and Granger (1987). This test is based on the residuals from the cointegrating regression. The cointegrating regression for the i th supplier is the regression of p_{it} on a constant and the p_{jt} ($j \neq i$). The intuition behind this test is that, if the series are cointegrated, then the errors from this regression should be stationary. It is implemented via a Dickey-Fuller test [in the form of (17) given above] on the residuals from the cointegrating regression for p_{it} . The null hypothesis of a unit root in the residual process corresponds to noncointegration, and the alternative of stationarity of the residual process

TABLE 3—REGRESSION-BASED TESTS FOR COINTEGRATION

Country	Augmented Dickey-Fuller statistic
China	-6.0708
Soviet Union	-4.0123
United States	-5.6352
South Africa	-5.7883
Australia	-3.4231

Notes: Critical values are -4.80 (0.01 level) and -4.15 (0.05 level).

corresponds to cointegration of the price processes. Table 3 contains these test statistics and their critical values. As can be seen from the table, the ADF tests on the residuals from the cointegrating regressions for China, the United States, and South Africa imply that the null hypothesis that the series are noncointegrating is rejected at the 0.01 level in favor of the alternative that they are cointegrating. For the Australia and Soviet Union cointegrating regression residuals, we find that the null hypothesis of noncointegration cannot be rejected at the 0.05 level. The critical values for the ADF statistics are those for the case $n = 5$ from Table 3 of Engle and Byung S. Yoo (1987). The results of this set of tests provide significant evidence in favor of the hypothesis that the five price series are cointegrating. These results support the use of (12) to model each price series.⁵

⁵We also performed the test for cointegration derived by Soren Johansen (1988). This test yielded a similar finding of cointegration.

TABLE 4—FIRST-ROUND ESTIMATES OF PRICE PROCESSES

Country	Parameter estimates			Specification test statistics	
	c_i	γ_i	β_i	AR(1) errors	ARCH(2) errors
China	-0.0022516 (0.0012798)	-0.71658 (0.16424)	0.34724 (0.11850)	-0.520287	0.01627
Soviet Union	-0.0026191 (0.0015359)	-0.74900 (0.18776)	-0.14908 (0.13111)	0.137535	2.4991
United States	-0.0035828 (0.0019120)	-1.16960 (0.24895)	-0.04572 (0.14853)	-0.821971	1.9785
South Africa	-0.0028731 (0.0012169)	-0.68449 (0.25212)	0.104030 (0.16425)	-0.638958	4.4891
Australia	-0.0032582 (0.0010532)	-0.12764 (0.12171)	-0.34435 (0.14205)	0.617483	1.9756

Note: Ordinary least-squares standard-error estimates are in parentheses below the coefficient estimates.

For each first-differenced price series, the model given in (12) with a constant term, z_{it-1} , and Δp_{it-1} was sufficient to represent adequately the behavior of each of the price processes over the sample and still not reject white-noise errors. Table 4 contains the results of these regressions. As expected, the signs of all of the parameters associated with the z_{it-1} are negative. Because of the presence of a lagged dependent variable in combination with z_{it-1} in the regression, the usual univariate Box-Pierce statistic for autocorrelation is not valid; instead, the auxiliary regression form of James M. Durbin's (1970) Lagrange multiplier (LM) test for AR(1) disturbances was computed. These statistics are asymptotically normal under the null hypothesis. For all of the models, there is very little evidence for this alternative against the null hypothesis of univariate white-noise errors. Generalizations of this LM test given in Trevor S. Breusch and Adrian R. Pagan (1980) against general fourth-order AR and MA processes were also performed, but the null hypothesis of white-noise errors could not be rejected for these cases either. As a test of the null hypothesis of a constant conditional variance of each of the price processes over time, we computed the LM test for ARCH errors derived by Engle (1982). This test statistic is computed as TR^2 from a regression of the squared residuals on lagged values of the squared residuals and a constant.

This statistic is asymptotically distributed as $\chi^2_{[k]}$ with k equal to the number of lagged residuals included in the auxiliary regression. These test statistics for the case of two lagged residuals are reported in Table 4. All of these statistics are considerably less than 5.991, the $\alpha = 0.05$ critical value from a $\chi^2_{[2]}$ random variable. Similar test results were obtained for the cases $k = 1$ and $k = 3$. The standard-error estimates for the coefficients reported in Table 4 are the usual single-equation ordinary least-squares (OLS) estimates, and as such do not take into account any of the restrictions of our structural model or the contemporaneous correlation between the ξ_{it} in the five price equations.

Conditional on these first-round estimates of the parameters of the price process, we then estimate λ and Ω by ML. Starting from the \sqrt{T} -consistent estimates of λ and Ω and the OLS estimates in Table 4, we then compute fully efficient ML estimates which impose the cross-equation restrictions implied by our risk-diversification model and the contemporaneous covariances between the ξ_{it} 's. Table 5 contains the converged ML estimate of Σ , the conditional covariance matrix of the price process. All ML parameter estimates were within two standard errors (using the consistent standard-error estimates computed from the converged ML parameter estimates as described in footnote 3) of the first-round set of consistent estimates of Γ ,

TABLE 5—MAXIMUM-LIKELIHOOD ESTIMATE OF $\Sigma \times 10^4$

Country	Country				
	China	Soviet Union	United States	South Africa	Australia
China	3.0	1.2	0.6	1.2	-0.08
Soviet Union	1.2	10.6	-3.2	2.2	0.4
United States	0.6	-3.2	2.6	-0.3	-0.2
South Africa	1.2	2.2	-0.3	1.3	0.3
Australia	-0.08	0.4	-0.2	0.3	1.0

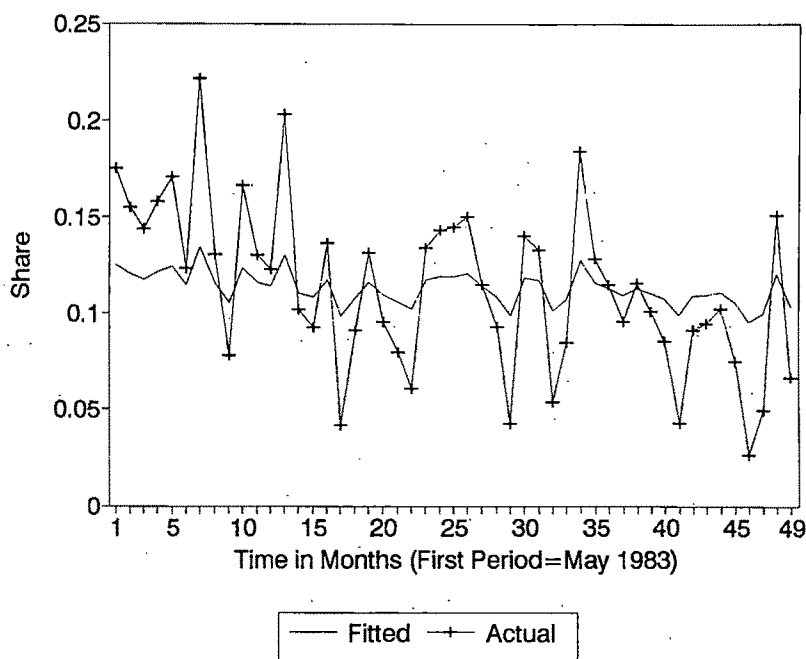


FIGURE 1. ACTUAL AND FITTED IMPORT QUANTITY SHARES FOR THE UNITED STATES FROM ML MODEL ESTIMATION

Σ , λ , and Ω . This result lends some credence to our ML estimation procedure, which jointly estimates Γ , Σ , λ , and Ω , and imposes the cross-equation restrictions implied by our structural model. These ML parameter estimates and associated consistent standard-error estimates allow an examination of the validity of the risk-diversification approach to input demand.

Figures 1 and 2 contain plots of the fitted versus actual values of the prices and shares from our ML model-estimation procedure

for the United States.⁶ The corresponding plots for the other four countries showed the same qualitative features as these plots and were therefore omitted to save space. The units on prices are thousands of yen per metric ton. The first thing to note from

⁶We are very grateful to a referee for suggesting many of the diagnostic tests and exploratory procedures discussed in the remainder of this section.

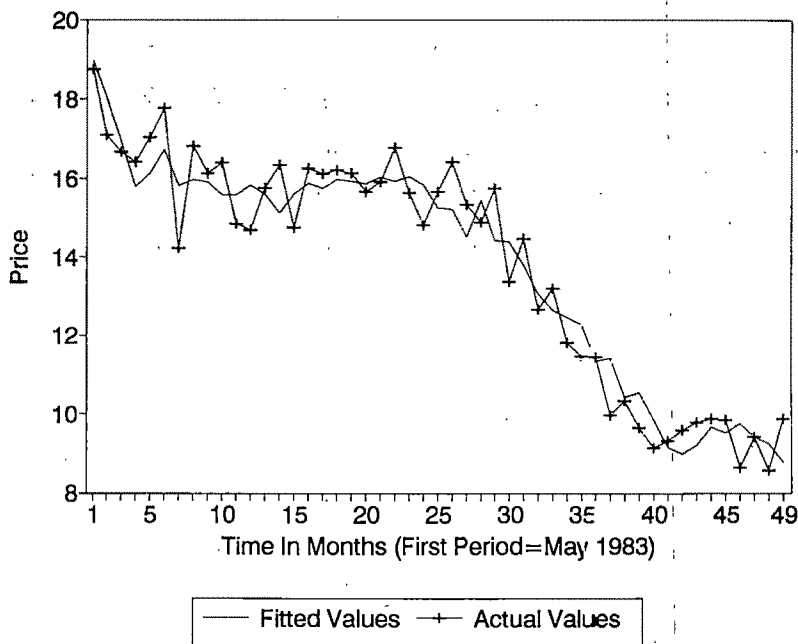


FIGURE 2. ACTUAL AND FITTED IMPORT PRICES IN THOUSANDS OF YEN PER METRIC TON OF STEAM COAL FOR THE UNITED STATES FROM ML MODEL ESTIMATION

these figures is that fitted prices follow actual prices much more closely than fitted shares follow actual shares. Nevertheless, the model fits well enough not to predict negative shares for any observations in our sample, despite the fact that the actual shares, especially those for the Soviet Union, are extremely close to zero. In order to give some idea of the explanatory power of our model we computed an " R^2 " for each of the price functions and share equations estimated. Analogous to the way that it is defined for the linear-regression model, we define $R^2 \equiv 1 - (\text{RSS}/\text{TSS})$, for

$$\text{RSS} = \sum_{t=1}^T (y_t - \hat{y}_t)^2$$

and

$$\text{TSS} = \sum_{t=1}^T (y_t - \bar{y})^2$$

where y_t is the dependent variable of the

TABLE 6— R^2 FOR PRICE AND SHARE EQUATIONS

Country	R^2	
	Price	Share
China	0.9527	0.3765
Soviet Union	0.8352	0.3501
United States	0.9339	0.3550
South Africa	0.9562	0.3632
Australia	0.9720	0.3642

equation, \hat{y}_t is the fitted value of y_t , and \bar{y} is the sample mean of y_t . In Table 6 we present these magnitudes for each price and share equation. Table 6 confirms the superior fit of the price equations. For the remainder of this section, we attempt to examine the validity of less restrictive models that include additional parameters or variables. We are unable to find substantial evidence against our structural model. Consequently, based on the specification tests discussed below, we conclude that this rather large unexplained variability in the

import shares is due either to factors that we have been unable to measure (which therefore cannot be incorporated into our model) or simply to nonsystematic deviations from optimizing behavior. We have modeled both of these phenomena by ϵ_t in (11).

We now discuss the tests of the restrictions implied by our structural model. All of them focus on the share equations, because our risk-diversification model implies restrictions on the parameters of these equations and on which variables enter these equations but no restrictions on the model estimated for the price process. These tests involve either testing the overidentifying restrictions of our model or testing for left-out variables. First we consider the tests for left-out variables. Because we are dealing with a time-series of import shares, one indication that we may have left out some important variable would be that the residuals from the share equations exhibit autocorrelation. To test for this we computed the Box-Pierce statistics for first-order autocorrelation for each of the residual vectors from the share equations. Under the null hypothesis of serially uncorrelated errors, these statistics are asymptotically distributed as $\chi^2_{[1]}$ random variables. As shown in Breusch and Pagan (1980), because there are no lagged dependent variables in the share equations, the Box-Pierce statistics are asymptotically equivalent to an LM test against AR or MA serial correlation of the order of the Box-Pierce statistic. These statistics are reported in Table 7. Tests for higher-order serial correlation yielded similar results.

Another potential left-out variable in the share equations is the lagged value of the share in each of the share equations. If there are rigidities in import shares due to the contractual nature of steam-coal purchases, one might expect some sort of model involving partial adjustment to the optimal share vector over time. As a consequence, lagged shares would help to explain the current values of the shares. To test this hypothesis we reestimated our complete model with lagged values of the dependent variable in each of the share equations. This

TABLE 7—TESTS FOR AUTOCORRELATION
IN SHARE EQUATIONS

Country	Test statistic
China	0.85
Soviet Union	2.33
United States	3.10
South Africa	0.56
Australia	2.01

Note: Critical value = 3.841 for $\alpha = 0.05$.

entails adding four parameters to our model. Recall that summability of the shares requires us to drop one share equation in the estimation. The likelihood-ratio statistic against this alternative hypothesis, computed as twice the difference between the log-likelihood functions for the unrestricted and restricted models, is asymptotically distributed as a $\chi^2_{[4]}$ random variable. This test statistic is equal to 8.22, which is less than 9.488, the $\alpha = 0.05$ critical value for a $\chi^2_{[4]}$ test. Based on the results of this test, lagged shares do not add any statistically significant explanatory power to our structural model. These results are consistent with the view that, although a portion of each period's purchases are made on long-term contracts, a sufficient amount of coal is also bought on the spot market so that these long-term contracts do not impose binding constraints on MITI's ability to adjust its import shares to what it believes is optimal for that period.

Our structural model also implies various cross-equation restrictions between and within the price and share equations. These cross-equation restrictions require that the parameters of Σ , the conditional covariance matrix of \mathbf{p}_t , enter into the share equations in a very specific way. We now consider two tests of the validity of these cross-equation restrictions. First rewrite (10) as

$$(19) \quad \mathbf{w}_t = \frac{\Sigma^{-1} \mathbf{t}}{\mathbf{t}' \Sigma^{-1} \mathbf{t}} + \phi \left[\frac{\Sigma^{-1} \mathbf{t}' \Sigma^{-1}}{\mathbf{t}' \Sigma^{-1} \mathbf{t}} - \Sigma^{-1} \right] \boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t.$$

The unrestricted form of this set of equations is

$$(20) \quad \mathbf{w}_t = \mathbf{A} + \mathbf{B}\boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t$$

where \mathbf{A} is an $n \times 1$ vector and \mathbf{B} is an $n \times n$ matrix. If we impose only summability ($\mathbf{1}'\mathbf{w}_t = 1$) on the vector of share equations, then this implies $\mathbf{1}'\mathbf{A} = 1$ and $\mathbf{1}'\mathbf{B} = 0$. After imposing these restrictions, we are left with 24 free parameters in \mathbf{A} and \mathbf{B} to estimate. Consequently, estimating (20) jointly with the price equations involves estimating 23 more parameters than does estimating (19) and the price equations, because ϕ is the only parameter estimated in (19) that is not also estimated in the price equations. Call this unrestricted model, which assumes only summability in the share equations, model S. Another set of restrictions is implied by the fact that $\boldsymbol{\Sigma}$ is a symmetric, positive definite matrix. By inspection of (19), one restriction on the structural model is that \mathbf{B} is symmetric. Combining the assumption that \mathbf{B} is symmetric with the summability assumption gives the restriction that the input-share equations are homogenous of degree zero in expected prices. If we impose symmetry and summability, this leaves 14 free parameters in \mathbf{A} and \mathbf{B} to estimate in equation (20). By the same logic as given above, moving from (20) with symmetry and summability to (19) requires imposing 13 nonlinear cross-equation constraints on the parameters of \mathbf{A} and \mathbf{B} . Call this more restricted model, which assumes summability and symmetry, model S&S. We also consider an additional restricted form of (20), which is not explicitly nested within (19) but embodies an interesting testable restriction about the impact of $\boldsymbol{\mu}_t$ on \mathbf{w}_t . This model assumes $\mathbf{B} = \mathbf{0}$, so that under this hypothesis, the conditional mean $\boldsymbol{\mu}_t$ is assumed to have no explanatory power in predicting import shares. Call this model BZ. For the model BZ, the import shares are assumed to be independent identically distributed draws rather than independent draws from a distribution with a conditional mean depending on $\boldsymbol{\mu}_t$.

The likelihood-ratio statistic testing for our complete structural model (15) against

the less restricted model S&S is equal to 20.02. This statistic, which is asymptotically distributed as a χ^2 random variable with 13 degrees of freedom, is less than 22.36, the $\alpha = 0.05$ critical value for the test. The likelihood-ratio test for (15) against the unrestricted model S is equal to 27.6. This statistic is asymptotically distributed as $\chi^2_{[23]}$. This statistic is less than the $\alpha = 0.05$ critical value of 35.17. The results of these hypothesis tests provide little, if any, evidence against the validity of the cross-equation restrictions implied by our structural model. Table 8 presents the ML estimates of \mathbf{A} and \mathbf{B} under the following three hypotheses: (i) the structural model given in (19) with the price-vector model in (14), (ii) model S with the price model (14), and (iii) model S&S with the price model (14). Although very few of the individual elements of \mathbf{A} and \mathbf{B} are precisely estimated, none of these parameter estimates is wildly inconsistent across the three models. As a general rule, the more restricted parameter estimate is always contained within two standard errors of the less restricted parameter estimate to its right. In other words, the approximate 95-percent confidence interval of the less restricted parameter estimate contains the corresponding more restricted parameter estimate. For example, in the case of B_{21} , the estimate from model S&S is 0.4219, which lies within approximately one standard error (0.7458) of the estimate of B_{21} from model S, 1.0158. In addition, the estimate of this parameter from model (19), 0.0731, lies within two standard errors (2.0×0.1775) of its estimate from model S&S.

To investigate the explanatory power of $\boldsymbol{\mu}_t$ for \mathbf{w}_t , we test the null hypothesis that $\mathbf{B} = \mathbf{0}$ against both S and S&S. The test statistic for BZ versus S&S is 31.81. Setting $\mathbf{B} = \mathbf{0}$ under S&S imposes 10 restrictions, so that the $\alpha = 0.01$ critical value for this hypothesis is 23.21. Consequently, BZ can be rejected against S&S at the 0.01 level of significance. The test statistic for BZ versus S is 39.39. Imposing $\mathbf{B} = \mathbf{0}$ relative to only summability of \mathbf{B} imposes 20 restrictions. The $\alpha = 0.01$ critical value for this test is 37.56, so the null hypothesis of BZ can be rejected against the alternative of model S.

TABLE 8—THREE MAXIMUM-LIKELIHOOD ESTIMATES OF A AND B FROM EQUATION (20)

Parameter ^a	Structural model (19) ^b	Model S&S from (20)	Model S from (20)	Model S from (20) ^c
A_1	0.0852 (0.0057)	0.0632 (0.0103)	0.0587 (0.0192)	
A_2	0.1189 (0.0087)	0.0883 (0.0211)	0.0661 (0.0451)	
A_3	0.2138 (0.0081)	0.1995 (0.0206)	0.1358 (0.0470)	
A_4	0.5505 (0.0122)	0.6193 (0.0233)	0.7162 (0.0740)	
B_{11}	-0.3594 (0.0749)	-0.6741 (0.2237)	-0.6174 (0.2275)	-0.1812 (0.3182)
B_{21}	0.0731 (0.0336)	0.4219 (0.1775)	1.0158 (0.7458)	0.2746 (0.2557)
B_{31}	0.2535 (0.0667)	0.0774 (0.3621)	-1.3188 (1.0319)	0.2192 (0.4255)
B_{41}	-0.0255 (0.0411)	0.1865 (0.2525)	0.9952 (0.8234)	0.3530 (0.2998)
B_{22}	-0.1019 (0.0208)	0.6619 (0.3054)	-2.1411 (2.3594)	-0.0030 (0.5930)
B_{32}	0.0002 (0.0340)	0.0687 (0.3222)	-0.0429 (0.6176)	0.2954 (0.7558)
B_{42}	0.0503 (0.0211)	-0.8268 (0.2953)	0.0498 (0.7847)	1.8168 (1.1021)
B_{33}	-0.3901 (0.0817)	0.8796 (1.1692)	3.5347 (1.9835)	-0.3492 (0.7651)
B_{43}	0.1043 (0.0431)	-0.6511 (0.5553)	-1.8740 (1.3807)	-1.3065 (1.1040)
B_{44}	-0.1753 (0.0359)	0.4971 (0.5344)	-0.6124 (0.6547)	0.3695 (1.0714)

Note: Estimated standard errors are shown in parentheses below the parameter estimates.

^aThe following definitions match the numbers with countries: China = 1, United States = 2, South Africa = 3, Australia = 4, and Soviet Union = 5. Some of the elements of A and B corresponding to the Soviet Union are not reported, because these estimates can be obtained from the summability restrictions.

^bThese are the estimates of A and B implied by the ML estimates of ϕ and Σ which arise from maximizing (16).

^cThe parameters in this column are remaining elements of B. They are listed in the following order from top to bottom: B_{15} , B_{12} , B_{13} , B_{14} , B_{25} , B_{23} , B_{24} , B_{35} , B_{34} , B_{45} . This order was selected to match the above-diagonal elements of B with their corresponding below-diagonal elements.

These two tests lead to the following weak conclusion: although the explanatory power of μ_i for w_i based on the R^2 criterion given in Table 6 is quite limited, it is statistically significantly different from no effect of μ_i on w_i .

One further test of our structural model attempts to address the question of whether or not λ , which determines the marginal rate of substitution between risk and return, is constant across the share equations. In

other words, is there a single rate at which expected cost is traded off against cost variability, independent of its source? To examine the hypothesis of a single marginal rate of substitution of risk for cost, we estimated our model subject to the summability restriction but allowing λ to vary across the share equations. Moving from this model to the model with a single λ involves imposing three restrictions, so that the likelihood-ratio test against this alternative hypothesis

TABLE 9—MAXIMUM-LIKELIHOOD ESTIMATE OF λ

Statistic	Value
$\hat{\lambda} \times 10^{-3}$	1.123
Standard error of $\hat{\lambda} \times 10^{-3}$	0.152
<i>t</i> statistic	7.365

would be asymptotically $\chi^2_{[3]}$ under the null hypothesis. This test statistic is 10.9, which lies above the $\alpha = 0.05$ critical value of 7.82 but below the $\alpha = 0.01$ critical value of 11.34. Consequently, we conclude that there seems to be some evidence against the hypothesis of a single MRS between risk and cost, but it is not overwhelming.

Having provided evidence in favor of our structural model as a reasonable summary of the observed data, we now turn to the question of the validity of the specific behavioral model giving rise to our structural model: the risk-diversification model of input choice. Table 9 contains the ML estimate of λ and its standard error.⁷ By substituting equation (8) into the log-likelihood function (16), we can see that the log-likelihood function is not defined at the point $\lambda = 0$. As a consequence of this, we cannot compute the distribution of $\hat{\lambda}$ under the hypothesis that $\lambda = 0$. However, for $\lambda = \varepsilon > 0$, the standard conditions necessary to test $H: \lambda = \varepsilon$ versus $K: \lambda > \varepsilon$ are satisfied. The data in Table 9 allow rejection (at a 0.05 level of significance) of the null hypothesis $\lambda = \varepsilon$ in favor of the alternative $\lambda > \varepsilon$ for all $872.96 > \varepsilon > 0$, where 872.96 is the solution in ε of $1.64 = (\hat{\lambda} - \varepsilon)/\text{SE}(\hat{\lambda})$, so that in this limited sense we can say that λ is significantly different from zero. Reparameterizing our model in terms of ϕ as given in (10), we obtain a likelihood function that is defined for all $\phi \in R$. Viewing our problem in this manner allows us to address the opposite question of whether or not Japan places a nonzero weight on expected cost in deter-

mining its optimal import mix. Applying the δ method to compute the asymptotic standard error of $\hat{\phi}$, we find that the null hypothesis that $\phi = 0$ can be rejected in favor of the alternative that ϕ is positive. These two views of our risk-diversification model allow us to conclude that, within the context of our structural model of input choice, Japan appears to attach a positive weight to both the conditional mean and conditional variance of total input cost in choosing its optimal supplier mix.

An informal but potentially informative check of our risk-diversification hypothesis is to compare the predictions of our model with those of the expected-cost-minimization model in terms of the deviations from the actual shares. Because the full maximum-likelihood function in (16) will tend to produce estimates of Γ and Σ that favor the risk-diversification model over the expected-cost-minimization model, we use the first-round estimates of the parameters of the price process (which do not impose any of the restrictions of our structural model) to perform this analysis. For comparison, we then present the results for final maximum-likelihood estimates.

The mechanics of our procedure are as follows. Our metric for comparison is the expected cost of the import mix purchased under each model. For the expected-cost-minimization model, for each time period t , this expected cost is computed by selecting the supplier with the smallest μ_{it} arising from our model for the price process and multiplying this expected price by the total quantity of coal delivered. This is precisely the expected cost for an agent using the expected-cost-minimization rule. Call this magnitude C'_{mc} . We then compute the difference between this expected cost and the expected cost of the actual bundle of imports purchased. Call this magnitude C'_{act} . Mathematically these expected costs are

$$C'_{mc} = \left(\min_i \mu_{it} \right) Q_t$$

$$C'_{act} = \mu'_t q_t$$

For the risk-diversification model, we compute the minimum-cost quantity mix that

⁷Other maximum-likelihood parameter estimates are not reported because of their agreement with (modulo two standard errors) the first-round estimates in Table 4.

yields the same conditional variance of cost as the actual quantities purchased. Define $V_t = \mathbf{q}_t' \Sigma \mathbf{q}_t$ as the conditional variance of the actual quantity vector \mathbf{q}_t . For each time period t , this minimum-conditional-cost quantity mix is the solution to

$$(21) \quad \min_{\mathbf{q}} \mathbf{q}' \boldsymbol{\mu}_t$$

subject to $V_t = \mathbf{q}' \Sigma \mathbf{q}$ and $Q_t = \mathbf{q}' \mathbf{v}$.

If \mathbf{q}_t^v is the solution to (21), then C_{rd}^t , the minimum cost of a quantity mix with variance V_t , is equal to $\mathbf{q}_t^{v'} \boldsymbol{\mu}_t$. We compute $C_{act}^t - C_{mc}^t$ and $C_{act}^t - C_{rd}^t$ for all observations. Because the absolute magnitudes of these differences in costs provide little intuition, we instead focus on the ratio of these differences to the actual expected cost. The sample average of $(C_{act}^t - C_{mc}^t) / C_{act}^t$ is approximately 0.169. This means that, averaging over our sample, the expected total import cost associated with the minimum-expected-cost criterion is 16.9 percent below actual expected import costs. The sample average of $(C_{act}^t - C_{rd}^t) / C_{act}^t$ is 0.035, so that, averaging over our sample, the minimum-cost import bundle having the same variance as the actual bundle imported, has an expected cost that is 3.5 percent below actual import costs. If we compute these two magnitudes using the final (as opposed to first-round) estimates of Γ and Σ , then the two numbers become 16.7 percent and 1.0 percent, respectively. Although there is an improvement in the conformity of the data to the risk-diversification model as a result of imposing the cross-equation restrictions between the share and price equations in the full maximum-likelihood procedure, the divergence of actual expected costs from those predicted by the risk-diversification hypothesis are minor when compared to the divergence of actual expected costs from those predicted by the expected-cost-minimization hypothesis. Although, as discussed above, the risk-diversification hypothesis cannot be explicitly tested due to the fact that the likelihood function is not well-defined at the point $\lambda = 0$, the evidence presented here suggests that it is a

superior model to the expected-cost-minimization model for describing the Japanese steam-coal import market.

V. Implications of the Risk-Diversification Model of Input Demand

We now examine the empirical implications of this estimated model of short-run input demand under price uncertainty. We are concerned with three general questions. What is the size of the risk premium associated with steam coal imported to Japan? What relationships between the observed prices and shares does the risk-diversification model imply, and are these relationships consistent with the observed data? How well does this model of input choice explain the three time-series properties of supplier prices and shares of steam coal imported to Japan discussed in the Introduction?

We first consider the question of the size of the risk premium on imported coal. To derive this magnitude, consider the mean-price versus standard-error-of-price frontier, plotted in Figure 3. Such frontiers can be plotted for each $(\boldsymbol{\mu}_t, \Sigma)$ pair in our sample. Figure 3 is constructed using $\hat{\Sigma}$, the maximum-likelihood estimate of Σ and $\bar{\boldsymbol{\mu}}$, the sample mean of $\boldsymbol{\mu}_t(\hat{\Gamma}, I_t)$, as a representative value of $\boldsymbol{\mu}_t$, where $\hat{\Gamma}$ is the ML estimate of Γ . Define $P_{pt}(\mathbf{w}) = \mathbf{w}' \mathbf{p}_t = \sum_{i=1}^5 w_i p_{it}$ as the actual weighted average price of steam-coal imports at time t , where $\sum_{i=1}^5 w_i = 1$. Let $E(P_{pt}(\mathbf{w})) \equiv \mathbf{w}' \boldsymbol{\mu}_t$ equal the expectation conditional on I_t of $P_{pt}(\mathbf{w})$ and $\sigma^2(P_{pt}(\mathbf{w})) \equiv \mathbf{w}' \Sigma \mathbf{w}_t$ equal its variance conditional on I_t .⁸ The mean-standard-error frontier given in Figure 3 comprises the set of $(E(P_{pt}), \sigma(P_{pt}))$ pairs such that $\sigma(P_{pt}(\mathbf{w}))$ is minimized over \mathbf{w} subject to the constraints $\mathbf{v}' \mathbf{w} = 1$ and $E(P_{pt}(\mathbf{w})) = K$, where K is some positive constant. Once a value of λ is specified, the solution of (2) implies a point on the mean-standard-error frontier corresponding to the optimal input mix

⁸For the remainder of this section, all expectations and variances are conditional on I_t , the firm's information set at time t .

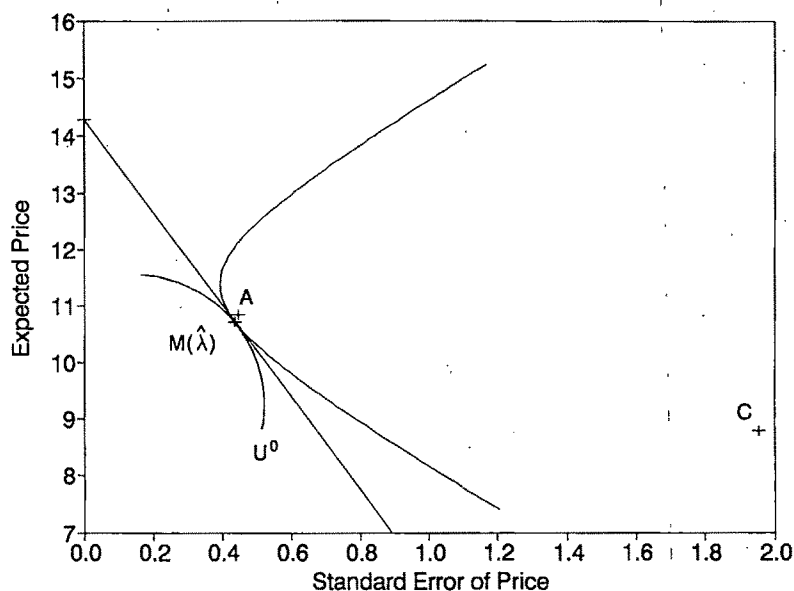


FIGURE 3. EFFICIENT FRONTIER COMPUTED USING THE ML ESTIMATE OF Σ AND THE SAMPLE AVERAGE OF ML ESTIMATES OF μ_t

for a given (μ_t, Σ) pair. The point labeled $M(\hat{\lambda})$ in Figure 3, is the optimal import mix (w_t^o) for $\hat{\lambda}$, the ML estimate of λ for $\mu_t = \bar{\mu}$ and $\Sigma = \hat{\Sigma}$. Because the efficient frontier depends on μ_t and Σ , the location of the point $M(\hat{\lambda})$ also depends on the values of these two magnitudes.

As an alternative measure of the relative fit of our model versus the expected-cost-minimization model, for $\hat{\Sigma}$ and $\bar{\mu}$, we also plot the expected-cost-minimizing bundle (subject to nonnegative input shares) and the sample average of the import bundles actually purchased. The expected-cost-minimizing point corresponds to purchasing only from the Soviet Union and is labeled point C in Figure 3. The point corresponding to the sample average of the import shares purchased is labeled A. A comparison of the distance between $M(\hat{\lambda})$ and A to the distance between A and C, provides further evidence against the expected-cost-minimization model versus the risk-diversification model.

The slope of the mean-standard-error frontier at the point $M(\hat{\lambda})$ is the rate at which Japan substitutes decreases in ex-

pected price for increases in the standard error of price. Define $P_t^o = w_t^{o'} p_t$ as the price corresponding to the point $M(\hat{\lambda})$. Where the tangent line to the point $M(\hat{\lambda})$ intersects the expected-price axis represents the expected price Japan would be willing to pay for riskless coal at time t , assuming that the marginal rate of substitution between risk and cost at $M(\hat{\lambda})$ [the point corresponding to $(E(P_t^o), \sigma(P_t^o))$ in Fig. 3] is constant for all levels of expected cost and standard errors of cost. We denote this price P_t^z because the portfolio of suppliers giving rise to this price is analogous to the zero-beta portfolio in the capital-asset pricing model (CAPM) with no riskless asset. We define the risk premium at time t (RP_t) as

$$(22) \quad RP_t = \frac{E(P_t^z) - E(P_t^o)}{E(P_t^o)}.$$

This risk premium has an alternative interpretation which can be understood without reference to the CAPM. The value of RP_t given in (22) is exactly equal to

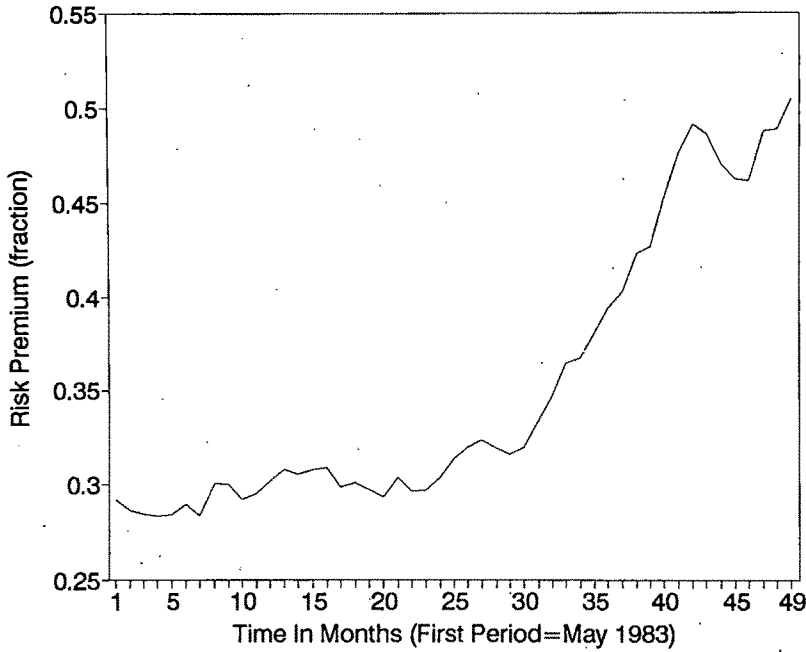


FIGURE 4. RISK PREMIUM (RP_t) FOR SAMPLE PERIOD

$-\varepsilon_{\mu_t^o, \sigma(P_t^o)}$, the negative of Japan's elasticity of import price risk with respect to expected import price. This elasticity is defined as

$$(23) \quad \varepsilon_{\mu_t^o, \sigma(P_t^o)} = \frac{d \log(\mu_t^o)}{d \log(\sigma(P_t^o))}$$

where

$$\mu_t^o = \mathbf{w}_t^{o'} \boldsymbol{\mu}_t \quad \text{and} \quad \sigma(P_t^o) = (\mathbf{w}_t^{o'} \boldsymbol{\Sigma} \mathbf{w}_t^o)^{1/2}$$

and where \mathbf{w}_t^o is defined in (7). In terms of the parameters of our structural model,

$$\varepsilon_{\mu_t^o, \sigma(P_t^o)} = - \frac{\sigma^2(P_t^o) \lambda}{\mu_t^o}.$$

Although $\varepsilon_{\mu_t^o, \sigma(P_t^o)} = -RP_t$, we still need to compute $E(P_t^z)$ in order to present other extensions of the analogy between the risk-diversification model and the CAPM.

An alternative methodology for computing P_t^z uses the intuition of the zero-beta CAPM. The price P_t^z arises from the portfolio of suppliers (weighted average price),

which has no market risk (market risk is defined as covariance with $P_t^o \equiv \mathbf{w}_t^{o'} \mathbf{p}_t$). To compute this portfolio, we solve for the minimum-variance weighted-average price subject to the constraint that its covariance with P_t^o is zero. The Lagrangian for this optimization problem takes the form

$$(24) \quad L = \frac{1}{2} \mathbf{w}_t^z' \boldsymbol{\Sigma} \mathbf{w}_t^z + \eta \left(\frac{1}{2} \mathbf{w}_t^z' \boldsymbol{\Sigma} \mathbf{w}_t^o \right) + \nu (1 - \mathbf{1}' \mathbf{w}_t^z)$$

where \mathbf{w}_t^z is the independent variable and η and ν are Lagrange multipliers associated with the constraints that the covariance of P_t^z with P_t^o is zero and that $\mathbf{1}' \mathbf{w}_t^z$ is equal to 1. The solution to (24) is

$$(25) \quad \mathbf{w}_t^z = \frac{\mathbf{w}_t^o - \mathbf{w}_t^{o'} \boldsymbol{\Sigma} \mathbf{w}_t^o (\boldsymbol{\Sigma}^{-1} \mathbf{1})}{1 - (\mathbf{w}_t^{o'} \boldsymbol{\Sigma} \mathbf{w}_t^o) (\mathbf{1}' \boldsymbol{\Sigma}^{-1} \mathbf{1})}.$$

Using (25) and our ML parameter estimates, we can compute $E(P_t^o) = \mu_t^o$ and $E(P_t^z)$ for all of our observations.

Figure 4 contains a time-series plot of RP_t based on the ML estimates of Γ , $\boldsymbol{\Sigma}$, and λ .

This risk premium ranges from 29 percent to 50 percent over the sample period, implying that Japan seems willing to pay 29–50 percent above the current market price for a supply of coal having no price risk. Recall that this calculation assumes that the marginal rate of substitution between risk and cost is constant over all risks and costs. If Japan's preferences for risk entail a declining marginal rate of substitution of risk for cost, then these numbers only represent an upper bound on the risk premium at time t . For some utility functions, they could be extremely conservative bounds on the risk premium. To illustrate this point, Figure 3 contains an indifference curve tangent to $M(\hat{\lambda})$ which exhibits a declining MRS.

By inspection of Figure 4, this risk premium exhibits an increasing time trend. A risk premium that increases with time is consistent with the view that, as Japan becomes more and more dependent on foreign sources of steam coal, as has been the case in recent years, the amount above the current weighted average market price Japan is willing to pay for coal with no price risk should increase. The interpretation of our results that is based on the elasticity of substitution between risk and cost [equation (23)] allows the following statement: at the point $M(\hat{\lambda})$, the loss in utility to Japan from a 1-percent increase in $\sigma(P_t^o)$ can be exactly offset by a 0.29–0.50-percent decrease in μ_t^o , depending on the time period in the sample.

We now turn to the issue of how well our risk-diversification model of input demand explains the time path of Japan's steam-coal import shares. To treat this issue, we first present one further implication of our model of input choice and discuss the applicability of this implication to our data. From equation (11), we know that the expected value of the observed vector of quantity shares is equal to the optimal vector of quantity shares based on μ_t , Σ , and λ . More precisely, the expectation of w_t is the vector of optimal import shares for μ_t , Σ , and λ , so that

$$E(w_t) = S(\mu_t, \Sigma, \lambda) = w_t^o.$$

This condition states that the expectation

of the actual import share Japan chooses is a point on the efficient $[E(P_{p_t}), \sigma(P_{p_t})]$ frontier. Using w_t^o , the optimal import-share vector for period t , we can construct a measure of risk for each supplier's price relative to $P_t^o = w_t^{o'} p_t$, analogous to the market-specific measure of risk for each security in the CAPM. For this reason, we denote the market-specific measure of risk for supplier i in period t by β_{it} and define it as

$$\beta_{it} = \frac{\text{Cov}(P_t^o, p_{it})}{\text{Var}(P_t^o)}$$

where p_{it} is supplier i 's price. The covariance and variance in the expression for β_{it} are conditional on I_t , the information set at time t . Consequently, because the composition of P_t^o will change each time period as μ_t changes, both the numerator and denominator of β_{it} will vary over time. Hence, β_{it} will also change over time. Figure 5 contains the plot of the β_{it} for all suppliers over the sample period. Recall, that, by construction, P_t^o has a β_t of 1 for all t , just as the β of the market portfolio in the CAPM is equal to 1.

Using logic similar to that used to derive the security market line in the CAPM, we can derive a relationship between the β_{it} and $E(p_{it}|I_t)$ as follows:

$$(26) \quad E(p_{it}|I_t) = E(P_t^z|I_t) + [E(P_t^o|I_t) - E(P_t^z|I_t)]\beta_{it}$$

where $E(P_t^o|I_t)$ is the conditional expectation of P_t^o and $E(P_t^z|I_t)$ is the conditional expectation of P_t^z . The derivation of this result exactly parallels the derivation of the zero-beta form of the security market line in the CAPM. The portfolio w_t^o is analogous to the market portfolio in the CAPM model. Thomas E. Copeland and J. Fred Weston (1983 pp. 198–200) provide a straightforward derivation of the zero-beta form of the security market line.

Note that relationship (26) depends on $E(P_t^o|I_t)$, a magnitude that explicitly depends on $S(\mu_t, \Sigma, \lambda) = w_t^o$, the optimal import shares derived from our risk-diversifi-

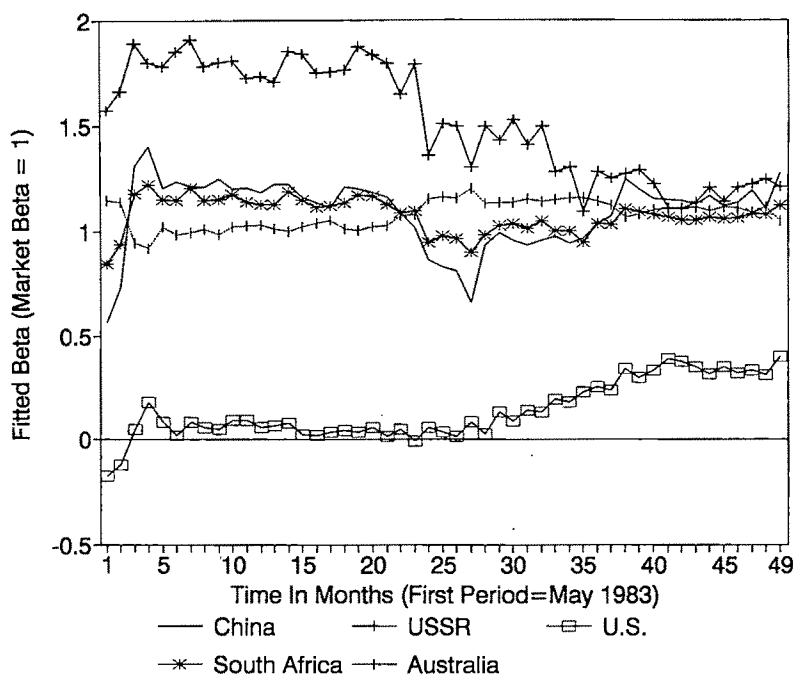


FIGURE 5. β 's OVER SAMPLE PERIOD, COMPUTED FROM EFFICIENT SHARES (β_{it})

cation model.⁹ Hence, another check of the reasonableness of our structural model is to recompute the β_{it} using w_t instead of w_t^0 to see whether or not there is a linear relationship between these β_{it} (call them β_{it}^a) and $E(p_{it}|I_t)$, as suggested by equation (26). Figure 6 contains a plot of the β_{it}^a . The levels and pattern of the β_{it}^a over time are quite similar to those followed by the β_{it} based on w_t^0 , but the time series of β_{it}^a is clearly more volatile than that of β_{it} . In Figure 7, we plot the sample average of the β_{it}^a against the sample mean of the μ_{it} . Although there are only five points on the plot, the relationship between the sample means of the μ_{it} and β_{it}^a is very well approximated by a straight line. Consequently, using the observed share data to construct the risk measures, a linear relation similar

to that in (26) seems to hold with $E(P_t^0|I_t)$ replaced by $E(P_t|I_t)$, where $P_t = w_t' p_t$.

We are now in a position to address the three puzzles presented in the Introduction. The first puzzle is why the United States remains in the market despite its consistently high price. Figure 5 shows that the United States consistently has the lowest market-specific measure of risk associated with it. In fact, β for the United States is negative in some periods, and for the most part it hovers around zero, indicating that United States coal is a good hedge against variations in the market price of coal. The very low market-specific risk associated with the price of U.S. coal explains why the United States services a sizeable share of the market even though its price path lies above those of the other four countries and has the second-largest conditional variance (see Table 5). Furthermore, the negative elements in the United States' row and column in Σ explains how this β close to zero comes about.

⁹We are grateful to a referee for suggesting the following procedure to examine validity of equation (26).

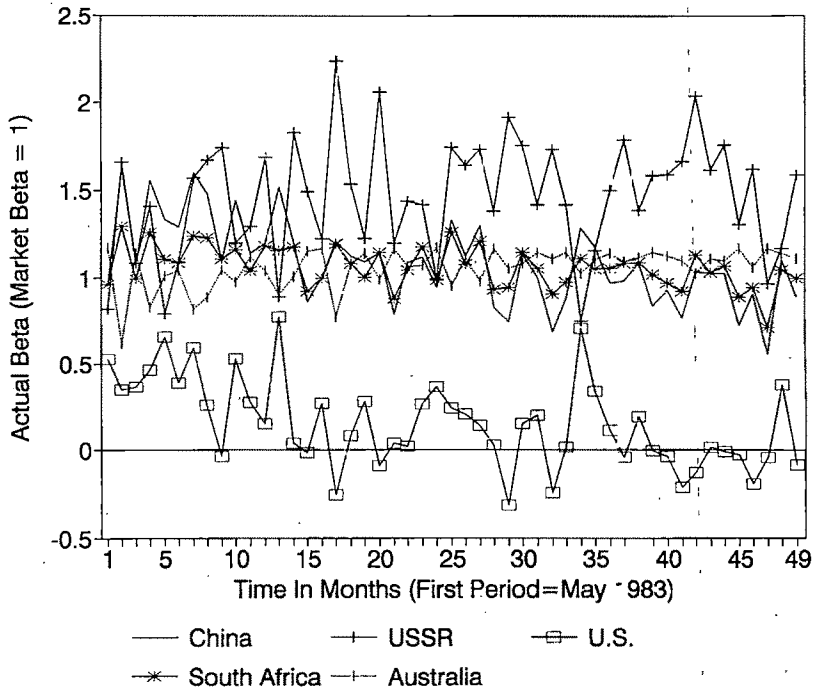


FIGURE 6. β 's OVER SAMPLE PERIOD, COMPUTED FROM ACTUAL SHARES (β_{it}^0)

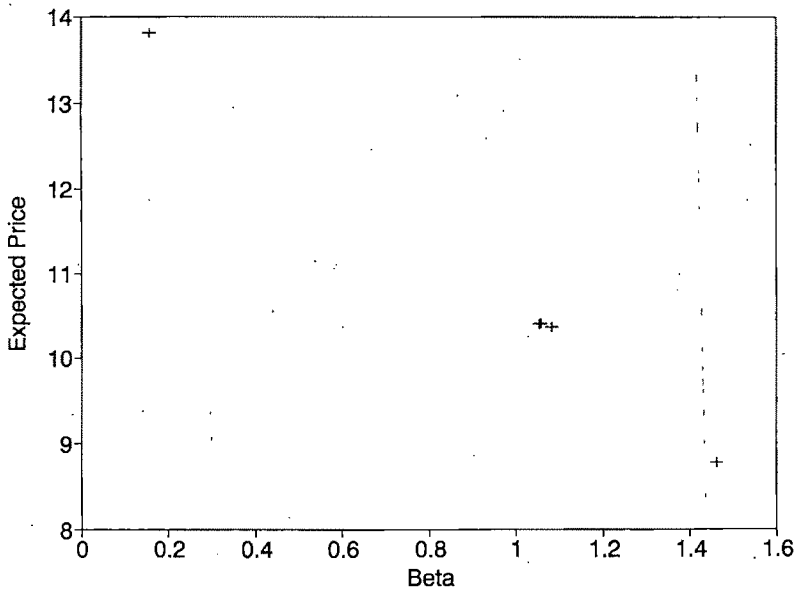


FIGURE 7. PLOT OF SAMPLE MEAN OF β_{it}^a VERSUS SAMPLE MEAN OF μ_{it}

The second puzzle is why the Soviet Union is consistently the cheapest supplier but never captures much of the market. Figure 5 also shows that the Soviet Union consistently has the highest market-specific measure of risk. In addition, the Soviet Union price has the highest conditional variance (see Table 5). These two risk measures illustrate why the Soviet Union has the smallest market share despite having the lowest price in almost all periods. From equation (26), we can see that the high level of market-specific risk associated with this supplier must be compensated for in terms of a low expected supply price in order for Japan to have nonzero demand for this coal.

The last puzzle concerns why South Africa and Australia have similar prices but very different market shares. This can be answered by inspection of our estimate of Σ in Table 5. Australia has the smallest conditional variance in price, and more importantly, its price has virtually no conditional covariance with any of the other prices. Both of these points imply that its market share should be substantially larger than that of South Africa, which has a higher conditional variance and higher conditional covariances with the other suppliers besides Australia. Finally, the similar time-series behavior of the β 's associated with Australia and South Africa explain, in part, why the two price processes from these countries are very similar and why the sample averages of the two price series are essentially the same.

VI. Conclusions and Policy Implications

The risk-diversification model of input demand seems to provide a useful framework for making economic sense of several puzzling anomalies in the Japanese steam-coal import market. Clearly, there are other models and factors that could explain the observed market shares; however, as mentioned earlier, the substantial anecdotal evidence for the applicability of the risk-diversification model of input demand makes an examination of its validity of particular interest and relevance.

The policy implications of these results for suppliers of Japanese steam-coal im-

ports fall into two broad categories. The first, perhaps more naive, view of these results is that because Japan seems to be willing to pay a premium for stable prices, a country interested in supplying more of its coal to Japan should attempt to stabilize its price of coal in yen to Japan. This view ignores the fact that much of the price uncertainty is due to factors beyond the control of coal suppliers. Supply interruptions, domestic price inflation in the country of origin, demurrage costs, exchange-rate fluctuations, and price inflation in Japan all affect the price of coal in yen to Japan. Consequently, perhaps a more sophisticated view of these results is that, as long as each supplier's price process has some component of its variation that is linearly independent of the variation in the prices of other suppliers, this supplier should have a nonzero market share whenever its prices are not too high above the prices of the other suppliers.

Perhaps the most significant result to come out of our paper is the development of a rigorous but implementable methodology for representing input demand under price uncertainty and for investigating the hypothesis of risk-diversification behavior in that framework. Future applications of this risk-diversification model of input demand are plentiful. Any industry in which a large portion of variable costs is taken up by a single homogeneous factor of production represents a potential test of the risk-diversification approach to input demand.

APPENDIX

This appendix describes the construction of the price and quantity series used in the empirical analysis. On a monthly basis, The Japan Tariff Association (JTA) publishes *Japan Exports and Imports, Commodity by Country*. This document gives the quantity (in metric tons) and the value (in thousands of yen) of imports, by country of origin, of various types of coal. During the sample period from May 1983 to May 1987, Japan imported coking, anthracite, lignite, and steam coal, which the JTA further decomposed into eight commodity classes. In terms of the Japanese Ministry of Finance's *Com-*

modity Classification for Foreign Trade Statistics classification system, steam coal is defined as commodity numbers 27.01-129 and 27.01-199. Consequently, to construct the total quantity and value of steam-coal imports from each country for each month, we computed the quantity and value totals for each country over these two commodity numbers. The price in thousands of yen per metric ton for a given country is obtained by dividing total value of shipments in the month by the total quantity of shipments in that month. If contacted at the address given in the initial footnote, the first author is willing to provide a machine-readable file containing these price and quantity data on a floppy diskette in DOS format.

REFERENCES

- Batra, Raveendra N. and Ullah, Aman, "Competitive Firm and the Theory of Input Demand under Price Uncertainty," *Journal of Political Economy*, June 1974, 82, 537-48.
- Berndt, Ernst K., Hall, Bronwyn H., Hall, Robert E. and Hausman, Jerry A., "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement*, October 1974, 3, 653-65.
- Blair, Roger D., "Random Input Prices and the Theory of the Firm," *Economic Inquiry*, June 1974, 12, 214-26.
- Breusch, Trevor S. and Pagan, Adrian R., "The Lagrange Multiplier Test and its Application to Model Specification in Econometrics," *Review of Economics Studies* January 1980 (Econometrics Issue), 47, 239-53.
- Copeland, Thomas E. and Weston, J. Fred, *Financial Theory and Corporate Policy*, Reading, MA: Addison-Wesley, 1983.
- Dickey, David A. and Fuller, Wayne A., "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, June 1979, 74, 427-31.
- Durbin, James M., "Testing for Serial Correlation in Least Squares Regression When Some of the Regressors Are Lagged Dependent Variables," *Econometrica*, May 1970, 38, 410-21.
- Engle, Robert F., "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, July 1982, 50, 987-1007.
- _____ and Granger, Clive W. J., "Co-integration and Error Correction: Representation, Estimation, and Testing," *Econometrica*, March 1987, 55, 251-76.
- _____ and Yeo, Byung S., "Forecasting and Testing in Co-integrated Systems," *Journal of Econometrics*, May 1987, 35, 143-60.
- Fuller, Wayne A., *Introduction to Statistical Time Series*, New York: Wiley, 1976.
- Johansen, Soren, "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, June/September 1988, 12, 231-54.
- Lehmann, Erich L., *Theory of Point Estimation*, New York: Wiley, 1983.
- Sandmo, Agnar, "Competitive Firm under Price Uncertainty," *American Economic Review*, March 1971, 61, 65-73.
- Stock, James H., "Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors," *Econometrica*, September 1987, 55, 1035-56.
- Tukenmez, Ercan and Tuck, Nancy, "Coal-Exporting Countries: The Asian Market," U.S. Department of Energy Report DOE/EIA-0462, Washington, DC: U.S. Government Printing Office, 1984.
- West, Kenneth D., "Asymptotic Normality, When Regressors Have a Unit Root," *Econometrica*, November 1988, 56, 1397-1418.
- Wu, Yuan-li, *Japan's Search for Oil*, Stanford, CA: Hoover Institution Press, 1977.
- Japan Export and Imports: Commodity by Country*, Tokyo: Japan Tariff Association, monthly, 1983-1987.
- Ministry of International Trade and Industry, *The MITI Handbook*, Tokyo: Japan Trade and Industry Publicity, Inc., 1979/1980.
- TEX Report, *1956 Coal Manual*, Tokyo: TEX Report, Ltd., 1986.
- United States Department of Energy, "Interim Report of the Interagency Coal Export Task Force," U.S. Department of Energy Report DOE/FE-0012, Washington, DC: U.S. Government Printing Office, 1981.

Sticky Prices As Coordination Failure

By LAURENCE BALL AND DAVID ROMER*

This paper links the "coordination failure" and "menu cost" approaches to the microeconomic foundations of Keynesian macroeconomics. If a firm's desired price is increasing in others' prices, then the gain from price adjustment after a nominal shock is greater if others adjust. This "strategic complementarity" leads to multiple equilibria in the degree of rigidity. Welfare may be much higher in the equilibria with less rigidity. Thus, nominal rigidity arises from a failure to coordinate price changes. (JEL E12, E30)

Keynesian macroeconomics waned in the 1970's because economists grew disenchanted with its weak microeconomic foundations. The central difficulty was that Keynesian models were based on ad hoc rigidities in nominal wages and prices. The 1980's produced two approaches to reviving Keynesian theory. The "menu cost" literature (N. Gregory Mankiw, 1985; George Akerlof and Janet Yellen, 1985) seeks to provide rigorous explanations for nominal rigidities. These papers argue that small frictions in price setting are enough to produce large nominal rigidities. In contrast, the "coordination failure" literature (Russell Cooper and Andrew John, 1988) abandons nominal rigidities and seeks alternative foundations for Keynesian models. The central idea is that many economic activities, such as production (e.g., John Bryant, 1983), trade (e.g., Peter Diamond, 1982), and investment (e.g., Nobuhiro Kiyotaki, 1988), exhibit "synergism" or "strategic complementarity": one agent's optimal level of activity depends positively on others' activity. Strategic complementarity can lead to multiple equilibria, with high-activity equilibria superior to low-activity equilibria.

Thus, an economy may be stuck in an "underemployment equilibrium" even though a superior equilibrium exists.

Models with nominal rigidities and models with coordination failures are often presented as competing paradigms.¹ This paper shows that this view is incorrect. We take a step toward unifying the foundations of Keynesian economics by showing that the two sets of ideas are highly complementary. Nominal rigidity arises from a failure to coordinate price changes. This failure has the essential features of coordination failures in previous models. Flexibility in one firm's price increases the incentives for other firms to make their prices flexible. This strategic complementarity leads to multiple equilibria in the degree of nominal rigidity. Equilibria with less rigidity (more active price adjustment) are often Pareto superior to equilibria with more rigidity.

These results contribute to our understanding both of coordination failure and of nominal rigidity. The range of Keynesian phenomena explained by coordination failures is greatly expanded. Previous coordination-failure models contain only real variables and thus ascribe no role to monetary policy or other determinants of nominal spending. Our results suggest that coordination failure is at the root of inefficient non-neutralities of money. Theories of nominal rigidities gain new empirical and policy im-

*Department of Economics, Princeton University, Princeton, NJ 08544; and Department of Economics, University of California, Berkeley, CA 94720. We are grateful for helpful comments from Ben Bernanke, Stephen Cecchetti, Andrew Policano, the referees, and participants in the May 1987 conference on coordination failure at the University of Iowa. The National Science Foundation provided financial support.

¹See, for example, Cooper and John (1988 pp. 441-2) and Diamond (1982 p. 881).

plications. As we argue below, strategic complementarity in price adjustment helps to explain variation in nominal rigidity across countries and over time. Our finding of coordination failure suggests a role for government intervention, either to improve coordination of price adjustment or to offset the effects of rigidity through active monetary policy.

We study the coordination of price adjustment in a model similar to those in Olivier Blanchard and Kiyotaki (1987) and Ball and Romer (1989a). Section I describes the model, and Section II presents our main results. In the model, imperfectly competitive firms decide whether to pay a small cost of adjusting prices after a nominal shock. Previous work shows that considerable rigidity can be an equilibrium even if it results in large, highly inefficient fluctuations in output. This paper shows that there are additional equilibria with less rigidity, and often with higher welfare. Specifically, for a range of realizations of the shock, both full adjustment of prices and complete non-adjustment are equilibria; this implies that an economy facing a distribution of shocks possesses a continuum of equilibrium degrees of rigidity. The size of the continuum is increasing in the degree of strategic complementarity in price adjustment.²

²The surveys of menu-cost models by Blanchard (1987) and Julio Rotemberg (1987) contain other discussions of multiple equilibria in the degree of rigidity. Rotemberg's argument is closer to ours. Costas Azariadis and Cooper (1985) and Roger Farmer and Michael Woodford (1984) present overlapping-generations models in which both flexible and sticky prices are equilibria. These models differ from ours both in the meaning of price rigidity and in the source of multiple equilibria. In overlapping-generations models, money serves only as a store of value. Thus, a sticky price level means sticky real asset prices. In our model, money is the medium of exchange, and so sticky nominal prices mean sticky transactions prices. In Azariadis and Cooper (1985) and in Farmer and Woodford (1984), the source of multiple equilibria is the more general fact that overlapping-generations models have a large indeterminacy of equilibria (Woodford, 1984); the flexible and sticky price equilibria are just two of many. In our model, multiple equilibria arise from the combination of menu costs and strategic complementarity in price-setting, and (with the minor exception discussed

Section III sketches two extensions of our analysis. First, we introduce heterogeneity among price setters, so that some prices adjust to a shock and others do not. There can be multiple equilibria in the proportion that adjust and, hence, in the size of the shock's real effects. Second, we consider a simple dynamic model in which firms choose between adjusting prices every period and every two periods. Here, there are multiple equilibria in the frequency of adjustment and, hence, in the dynamics of the price level. In addition, this example illustrates a difference between our model and other coordination-failure models: while the economy possesses multiple short-run equilibria, it converges to a unique long-run equilibrium.

Section IV concludes by discussing the model's empirical and policy implications.

I. The Model

The model is similar to the one in Ball and Romer (1989a), which is based on Blanchard and Kiyotaki (1987). While Blanchard and Kiyotaki specify both goods and labor markets, we assume for simplicity that the economy consists of "yeoman farmers" who sell goods produced with their own labor. That is, we suppress the labor market and focus on rigidities in output prices. Subsection A describes tastes and technology, and Subsection B describes how we measure nominal rigidity.

A. Tastes and Technology

There is a continuum of yeoman farmers indexed by i and distributed uniformly on $[0, 1]$. Each farmer produces a differentiated good, sells this product, and purchases the products of all other farmers. Farmers take each other's prices as given.

Farmer i 's utility function is

$$(1) \quad U_i = C_i - \frac{\varepsilon - 1}{\gamma \varepsilon} L_i^\gamma - z D_i$$

in footnote 7) complete rigidity and full adjustment are the only equilibrium responses to a shock.

where

$$(2) \quad C_i = \left[\int_{j=0}^1 C_{ij}^{(\varepsilon-1)/\varepsilon} dj \right]^{\varepsilon/(\varepsilon-1)}$$

and where L_i is farmer i 's labor supply; C_i is an index of farmer i 's consumption; C_{ij} is farmer i 's consumption of the product of farmer j ; z is a small positive constant (the menu cost); D_i is a dummy variable equal to 1 if farmer i changes his nominal price; ε is the elasticity of substitution between any two goods ($\varepsilon > 1$); and γ measures the extent of increasing marginal disutility of labor ($\gamma > 1$). The coefficient on L_i^γ in (1) is chosen for convenience. Finally, farmer i has a linear production function:

$$(3) \quad Y_i = L_i$$

where Y_i is farmer i 's output.

The utility function determines the demand for farmer i 's product, given aggregate consumption and the farmer's relative price:

$$(4) \quad Y_i^D = C \left(\frac{P_i}{P} \right)^{-\varepsilon}$$

where P_i is the price of good i , C is aggregate consumption, P is the aggregate price index, and

$$(5) \quad C = \int_{j=0}^1 C_j dj$$

$$(6) \quad P = \left[\int_{j=0}^1 P_j^{1-\varepsilon} dj \right]^{1/(1-\varepsilon)}$$

Farmer i 's consumption equals his real revenues:

$$(7) \quad C_i = \frac{P_i Y_i}{P}$$

[See Blanchard and Kiyotaki (1987) for derivations of (4)–(7).] Substituting (3), (4), and (7) into (1) yields farmer i 's utility as a function of aggregate consumption and his

relative price:

$$(8) \quad U_i = C \left(\frac{P_i}{P} \right)^{1-\varepsilon} - \frac{\varepsilon-1}{\gamma\varepsilon} C^\gamma \left(\frac{P_i}{P} \right)^{-\gamma\varepsilon} - zD_i.$$

To make nominal disturbances possible, we assume that money is required for transactions, so aggregate nominal spending equals the money stock:

$$(9) \quad PC = M.$$

Julio Rotemberg (1987) describes a specific transactions technology that gives rise to (9). Purchases must be made with money. At the start of the model's single period, a central bank distributes an amount of money M to farmers. A dollar can be spent only once during a period (velocity equals one), and farmers must repay the bank at the end of the period. These assumptions imply (9) and assure that the budget constraint (7) is satisfied.³

Substituting (9) into (8) yields

$$(10) \quad U_i = \left(\frac{M}{P} \right) \left(\frac{P_i}{P} \right)^{1-\varepsilon} - \frac{\varepsilon-1}{\gamma\varepsilon} \left(\frac{M}{P} \right)^\gamma \left(\frac{P_i}{P} \right)^{-\gamma\varepsilon} - zD_i.$$

³Two details of this story deserve mention. First, individuals choose how much money to receive from the bank. The bank equates the total demand for money to the supply, M , by adjusting the amount that agents are required to repay. The demand is infinite if the required repayment is less than one-for-one and zero if it is greater; thus, the equilibrium repayment is one-for-one. Second, aggregate nominal spending is $\int_{j=0}^1 P_j Y_j dj$. Using (7) and (5), this can be rewritten as $\int_{j=0}^1 P C_j dj = PC$. Thus, (9) follows from the assumption that aggregate spending equals M .

Obviously the ideas behind (9) are more general than Rotemberg's model. Blanchard and Kiyotaki derive (9) by putting money in the utility function. In a dynamic model, money can be introduced through a more realistic cash-in-advance constraint or through overlapping generations [although these models may not yield (9) exactly]. Our approach introduces money as simply as possible.

Differentiation of (10) shows that farmer i 's utility-maximizing price, neglecting the menu cost, is

$$(11) \quad P_i^* = P^\phi M^{1-\phi}$$

with

$$\phi = 1 - \frac{\gamma - 1}{\gamma\epsilon - \epsilon + 1} \quad 0 < \phi < 1$$

where ϕ is the elasticity of P_i^* with respect to the aggregate price level. Equation (11) implies that, in the absence of menu costs, symmetric equilibrium occurs when $P_i = P = M$. Finally, combining (10) and (11) yields farmer i 's utility as a function of real balances, the ratio of his price to the utility-maximizing level, and the menu cost:

$$(12) \quad U_i = \left(\frac{M}{P}\right)^{\gamma(1-\epsilon+\epsilon\phi)} \\ \times \left[\left(\frac{P_i}{P_i^*}\right)^{(1-\epsilon)} - \frac{\epsilon-1}{\gamma\epsilon} \left(\frac{P_i}{P_i^*}\right)^{-\gamma\epsilon} \right] - zD_i \\ = V\left(\frac{M}{P}, \frac{P_i}{P_i^*}\right) - zD_i.$$

The analysis below uses several properties of the function $V(M/P, P_i/P_i^*)$ in the vicinity of $M/P = 1$, $P_i/P_i^* = 1$, the equilibrium in the absence of menu costs. This function is increasing and concave in M/P : $V_1(M/P, P_i/P_i^*) > 0$, $V_{11}(M/P, P_i/P_i^*) < 0$ (subscripts denote partial derivatives). Intuitively, a rise in M/P benefits a farmer by raising aggregate consumption and thus shifting out the demand curve that he faces. Concavity means that farmers dislike fluctuations in demand. The source of this risk aversion is the increasing marginal disutility of producing output: $\gamma > 1$ in (1). Finally, since $P_i/P_i^* = 1$ maximizes utility by defi-

nition,⁴

$$V_2\left(\frac{M}{P}, 1\right) = 0 \\ V_{22}\left(\frac{M}{P}, 1\right) < 0 \quad \forall \quad \frac{M}{P}.$$

B. Nominal Rigidity

Here, we describe the basic experiment that we consider. The economy begins with $M = 1$ and $P_i = P_i^* = 1 \forall i$. One can think of this as the situation in an earlier period when the economy is at its long-run equilibrium. In the current period, a shock occurs: M changes to $1 + x$. Each farmer chooses between keeping his price at 1 or paying the menu cost and changing his price to the new P_i^* . We determine the circumstances under which adjustment and nonadjustment of prices are equilibria.

While natural, the assumption that prices initially equal 1 is ad hoc (the "earlier period" is not explicit). Therefore, in an Appendix we follow our earlier article (Ball and Romer, 1989a) in assuming that farmers choose initial prices optimally, given a distribution of shocks with mean 0. Farmers choose initial prices different from 1 (i.e., certainty equivalence fails), because utility is not quadratic. We find that the results in the text are altered only slightly.⁵

Our formulation assumes that farmers set prices in nominal terms and thus that they can eliminate the real effects of money only by adjusting their prices after observing M . A natural question is why farmers do not simply set indexed prices (i.e., announce a function relating their prices to M) and

⁴As this discussion suggests, our use of specific functional forms is not important for our main results. Aside from the properties of $V(\cdot)$ described here, the only essential assumption is strategic complementarity in utility-maximizing prices: P_i^* must be increasing in P .

⁵The less rigorous approach in the text is in a sense more realistic, since actual price rigidity is usually a failure to adjust from a previous price. This idea is captured rigorously in the dynamic model of Section III.

thereby eliminate the need for *ex post* adjustment.⁶ The answer is that indexing a price, like adjusting *ex post*, requires small amounts of effort. In this case, the "menu costs" include the costs of computing the number of dollars corresponding to indexed prices and of learning to think in real rather than nominal terms. As Bennett McCallum (1986) explains, it is easier to set prices in nominal terms—in units of money—because money is the medium of exchange. In other words, it is convenient to use the medium of exchange as the unit of account.

By assuming that if farmers achieve flexibility they do so through *ex post* adjustment, we are implicitly assuming that *ex post* adjustment is less expensive than indexation. Assuming the reverse does not change the basic character of our results. In this case, farmers choose between indexed prices and noncontingent prices before observing the monetary shock. They base their decision on the variance of the shock, which determines the expected cost of forgoing indexation. Just as our basic model produces multiple equilibria for a range of realizations of the shock, the alternative model produces multiple equilibria (indexation and nonindexation) for a range of variances.

II. Coordination Failure

This section presents our central results. Subsection A shows that both full adjustment of prices and complete nonadjustment are Nash equilibria for a range of sizes of the monetary shock. It follows that the economy possesses a continuum of equilibrium degrees of nominal rigidity. Subsection B compares welfare in the different equilibria.

A. Multiple Equilibria

We first determine when nonadjustment of all prices is an equilibrium. This is how

⁶In this model, indexation of individual prices to the aggregate price level would not accomplish the same thing. If each farmer set $P_i = P$, relative prices would be constant, but the aggregate price level (and hence real output) would be indeterminate.

previous menu-cost papers measure nominal rigidity. The condition for nonadjustment to be an equilibrium is that a representative farmer i chooses not to pay the menu cost if no other farmer pays. If farmer i maintains a rigid price along with the others, then $D_i = 0$. $P_i = P = 1$, which implies $M/P = M$, and using (11), $P_i/P_i^* = 1/M^{1-\phi}$. Thus, the farmer's utility is $V(M, 1/M^{1-\phi})$.

If farmer i pays the menu cost despite others' nonadjustment, then $D_i = 1$. Adjustment of one price does not affect the aggregate price level, so $P = 1$ and $M/P = M$. Adjustment allows farmer i to set $P_i = P_i^*$, so $P_i/P_i^* = 1$. Thus, farmer i 's utility is $V(M, 1) - z$.

These results imply that the representative farmer chooses not to pay the menu cost—and thus that rigidity is an equilibrium—if

$$(13) \quad G_N < z$$

$$G_N \equiv V(M, 1) - V\left(M, \frac{1}{M^{1-\phi}}\right).$$

G_N is the gain to a farmer from adjusting, given that others do not adjust. Rigidity is an equilibrium if G_N is less than the menu cost.

Taking a second-order approximation of G_N around $M = 1$ yields

$$\begin{aligned} (14) \quad G_N &\cong [V(1, 1) + V_1x + \frac{1}{2}V_{11}x^2] \\ &\quad - [V(1, 1) + V_1x + \frac{1}{2}V_{11}x^2 \\ &\quad + \frac{1}{2}V_{22}(1-\phi)^2x^2] \\ &= \frac{-(1-\phi)^2}{2}V_{22}x^2 \end{aligned}$$

where $x \equiv M - 1$ and where subscripts of V denote partial derivatives evaluated at $(1, 1)$ (recall that V_{22} is negative). The derivation uses the fact that $V_2(M/P, 1) = 0 \forall M/P$, which implies $V_2(1, 1) = V_{12}(1, 1) = 0$. Equation (14) shows that the gain from adjusting is increasing in the size of the shock. Equations (13) and (14) imply that the gain is less

than the menu cost, and so rigidity is an equilibrium, if $|x| < x_N$, where

$$(15) \quad x_N = \sqrt{\frac{-2z}{(1-\phi)^2 V_{22}}}$$

We now ask when price flexibility is an equilibrium. This occurs when farmer i chooses to pay the menu cost if all others pay. If all other farmers pay the menu cost, then $P = M$, so $M/P = 1$. If farmer i pays as well, then $D_i = 1$ and $P_i/P_i^* = 1$; thus, his utility is $V(1, 1) - z$. If farmer i does not pay the menu cost even though others do, then $D_i = 0$, $P_i = 1$, and, using (11), $P_i/P_i^* = 1/M$. In this case, farmer i 's utility is $V(1, 1/M)$.

These results show that farmer i pays the menu cost if

$$(16) \quad G_A > z$$

$$G_A \equiv V(1, 1) - V\left(1, \frac{1}{M}\right).$$

Flexibility is an equilibrium if G_A , the gain from adjusting given that others adjust as well, is greater than the menu cost. A second-order approximation yields

$$(17) \quad G_A = -\frac{1}{2}V_{22}x^2.$$

Like G_N , G_A is increasing in the size of the shock. Equations (16) and (17) imply that flexibility is an equilibrium if $|x| > x_A$, where

$$(18) \quad x_A = \sqrt{\frac{-2z}{V_{22}}}.$$

Combining (15) and (18) yields our central result:

$$(19) \quad \frac{x_N}{x_A} = \frac{1}{1-\phi}$$

in which ϕ , the elasticity of P_i^* with respect to P , is between 0 and 1. Thus, $x_A < x_N$. If

$|x|$ is between x_A and x_N , then both rigidity and flexibility are equilibria.^{7,8}

These results can be summarized as follows. For small monetary shocks ($|x| < x_A$), each farmer refuses to pay the menu cost regardless of others' decisions, and so rigidity is the only equilibrium. For large shocks ($|x| > x_N$), each farmer pays regardless of others, and so flexibility is the only equilibrium. However, for shocks of intermediate size ($x_A < |x| < x_N$), a farmer pays if and only if others do. The reason is that a farmer's gain from adjusting his price is greater if others adjust: G_A is greater than G_N . In Cooper and John's (1988) terminology, there is "strategic complementarity" in price adjustment. To see why, consider a positive shock for concreteness and recall that a farmer's utility-maximizing price, P_i^* , equals $P^\phi M^{1-\phi}$. If others keep their prices fixed at 1, P_i^* rises to $M^{1-\phi}$. However, if others adjust, P rises to M and P_i^* rises to $M > M^{1-\phi}$. That is, if others adjust, they change their prices in the same direction as the money supply, which pushes P_i^* farther from 1. Since the desired increase in P_i is larger, the incentive to adjust is larger.⁹

⁷One can show that, when both rigidity and flexibility are equilibria, there is a third equilibrium in which some farmers adjust and others do not and in which each farmer is indifferent about whether to adjust. This equilibrium is unstable: if slightly more than the required proportion of farmers adjust, then all farmers wish to adjust; if slightly fewer adjust, then none wishes to adjust.

⁸Our result that the model possesses multiple equilibria for some values of x does not appear to depend on our use of Taylor approximations. As explained below, the crucial condition for multiple equilibria is $G_A > G_N$. Without approximating, we are not able to show analytically that this holds for all parameter values, but extensive numerical calculations suggest that it does.

⁹Accommodating monetary policy would be another source of multiple equilibria. Suppose the money-supply rule is changed from $M = 1 + x$ to $M = 1 + c(P - 1) + x$, $0 < c < 1$. Since $P = 1$ if prices are rigid, x_N is not affected; but if prices are flexible, the equilibrium level of P and M is $1 + [x/(1-c)]$ rather than $1 + x$. As a result, $x_A = \{[-2z(1-c)^2]/V_{22}\}^{1/2}$ and $x_N/x_A = 1/[(1-\phi)(1-c)]$. Thus, accommodating monetary policy increases the range of multiple equilibria and makes multiple equilibria possible even if $\phi \leq 0$. Intuitively, accommodating policy creates an ad-

As this discussion suggests, strategic complementarity in price adjustment is tied to a simpler kind of strategic complementarity: the positive dependence of a farmer's utility-maximizing price in the absence of menu costs on the prices of others. The degree to which G_A exceeds G_N depends on ϕ , the elasticity of P_i^* with respect to P [see (14) and (17)].¹⁰ This implies that x_N/x_A is also increasing in ϕ [see (19)]. With strong strategic complementarity— ϕ close to 1—the range of multiple equilibria can be very large. Intuitively, changes in others' prices have a large effect on farmer i 's adjustment decision when they have a large effect on the farmer's desired price.¹¹

So far, our results concern equilibrium responses to a single shock. Now suppose that farmers face a distribution of shocks and choose rules for when to pay the menu cost. We restrict attention to equilibria in which all farmers pay the menu cost if $|x|$ is greater than a cutoff, x^* , that is, if the

money supply lies outside of $(1 - x^*, 1 + x^*)$. The cutoff x^* is a natural measure of the degree of rigidity. Our results imply that any value of x^* between x_A and x_N is an equilibrium; a farmer will adopt any value in this range as a cutoff if all others do. Thus, there is a continuum of equilibrium degrees of nominal rigidity.¹²

Finally, we note an unrealistic feature of our model: since complete adjustment of prices is a unique equilibrium when $|x| > x_N$, very large nominal shocks are necessarily neutral. In practice, large shocks appear to have large real effects; for example, sharp monetary contractions appear to cause deep recessions. Our result is an artifact of the simple static specification. As explained below, it disappears in dynamic versions of the model.

B. Welfare

Many coordination-failure models possess multiple equilibria that can be Pareto ranked. In particular, high-“effort” equilibria (for example, those with high levels of production) are often superior to low-effort equilibria. It is natural to ask whether this is the case in the current model. When there are multiple equilibria in the degree of price rigidity, is less rigidity (more effort expended on price adjustment) better?

To study welfare, we again assume that farmers face a distribution for the monetary shock, x , and pay the menu cost if $|x|$ exceeds a cutoff, x^* . For a symmetric distribution with mean zero, we derive the socially optimal value of x^* : the one that maximizes farmers' expected utility. To determine the welfare properties of equilibrium rigidity, we compare the optimal x^* to x_A and x_N , the endpoints of the range of equilibria. We continue to assume that farmers initially set their prices to 1, the equilibrium value in the absence of shocks; in the Appendix, we study the case in which

ditional source of strategic complementarity: when others raise their prices, M rises, which raises P_i^* .

¹⁰While the result that a farmer's utility-maximizing price increases with others' prices is clearly realistic, one can find cases in which it does not hold. For example, ϕ can be negative (prices can be strategic substitutes) if aggregate demand increases more than one-for-one with real money (as in Ball [1987]). If ϕ is negative, there is always a unique equilibrium in the fraction of farmers who adjust their prices.

¹¹In terms of the model, ϕ approaches 1 as γ approaches 1 (constant marginal utility of leisure) and as ϵ approaches infinity (a perfectly competitive product market). When γ approaches 1, if others do not change their prices, farmer i has no desire to change his: G_N approaches 0 and x_N approaches infinity. However, if others adjust, the benefits of adjusting with them are positive, and so the farmer adjusts if the shock is large enough: x_A approaches $\sqrt{2z/(\epsilon-1)}$. Thus, there are two equilibria for any

$$|x| > \sqrt{2z/(\epsilon-1)}.$$

On the other hand, if ϵ approaches infinity, then G_N approaches infinity and x_N approaches 0 (x_N/x_A still approaches infinity, because x_A approaches 0 more quickly than does x_N). When markets are competitive, a farmer's desired price change is small if others' prices are rigid, but the cost of forgoing even a small change is large. Formally, G_N approaches infinity because V_{22} grows more quickly than $(1-\phi)^2$ shrinks [see (14)].

¹²The economy also possesses equilibria with less natural rules for when to change prices. For example, the set of realizations of M for which prices are rigid can be an asymmetric range, $(1 - x_1^*, 1 + x_2^*)$, or even a disconnected set, $((M_1^*, M_2^*), (M_3^*, M_4^*))$.

initial prices are chosen optimally given the distribution of shocks.¹³

Recall that a farmer's utility is $V(1, 1) - z$ if all farmers pay the menu cost and $V(M, 1/M^{1-\phi})$ if none pays. Thus, since all pay if $|x| > x^*$, expected utility is

$$(20) \quad E[U_i] \\ = \{1 - [F(1 + x^*) - F(1 - x^*)]\} \\ \times [V(1, 1) - z] \\ + \int_{M=1-x^*}^{1+x^*} V\left(M, \frac{1}{M^{1-\phi}}\right) f(M) dM$$

where $F(\cdot)$ is the cumulative distribution function for M and $f(\cdot)$ is the density function. The first-order condition for the socially optimal x^* , denoted x_S , is

$$(21) \quad -2[V(1, 1) - z] \\ + V\left(1 + x_S, \frac{1}{(1 + x_S)^{1-\phi}}\right) \\ + V\left(1 - x_S, \frac{1}{(1 - x_S)^{1-\phi}}\right) = 0$$

where we use the fact that $f(1 + x) = f(1 - x)$ by our assumption that $f(\cdot)$ is symmetric around 1. A second-order approximation leads to

$$(22) \quad x_S = \sqrt{\frac{-2z}{V_{11} + (1 - \phi)^2 V_{22}}}$$

Our central welfare result follows from substituting the appropriate derivatives of

$V(\cdot)$ into (22) and the expressions for x_N and x_A :

$$(23) \quad x_A < x_S < x_N$$

Since $x_S < x_N$, there is a range of equilibrium values of x^* ($x_S < x^* < x_N$) with too much rigidity; in these equilibria, all farmers would be better off if the cutoff were lowered. Since $x_S > x_A$, there is a range of equilibria with too much flexibility. Finally, the social optimum ($x^* = x_S$) is always an equilibrium.

The reason that too much rigidity is possible is similar to the reason in Ball and Romer (1989a). Suppose that all farmers start with an arbitrary x^* . If one farmer lowers his cutoff while the others do not, the only benefit is that he sets $P_i = P_i^*$ more frequently; but if *all* farmers reduce x^* , there is an additional benefit. All prices adjust more frequently, and so the aggregate price level becomes more flexible. This reduces fluctuations in the real money stock and thus stabilizes the demand curves that farmers face. As explained above, farmers prefer stable demand because the disutility of labor is convex. Since the incentive for an individual to reduce x^* is smaller than the gain if all do, values of x^* above x_S can be equilibria.

Values of x^* below x_S can be equilibria (i.e., there can be too much flexibility) because a farmer's gain from raising x^* is also smaller if he does so by himself than if all do. If the others do not join the farmer in raising x^* , then for some shocks he does not adjust his price but others do. Others' adjustment increases movements in P_i^* , which raises the farmer's loss from nonadjustment. (Others' adjustment still benefits the farmer by stabilizing demand, but this effect is smaller.)

While both excessive rigidity and excessive flexibility are possible, the magnitudes of the losses are very different. Neglecting the menu cost, full flexibility is optimal ($x_S = 0$ when $z = 0$). Thus, the net loss from too much flexibility is bounded by the menu cost, which realistically is small. In contrast, Ball and Romer (1989a) show that the loss from too much rigidity can be arbitrarily

¹³We study average welfare given a distribution of shocks because the welfare effect of rigidity after an individual shock depends on the sign of the shock (Mankiw, 1985; Ball and Romer, 1989a). Nonadjustment to a fall in the money supply reduces output and welfare. However, nonadjustment to a positive shock increases output. This raises welfare because, under imperfect competition, the no-shock level of output is too low.

large. Intuitively, the private incentive to reduce x^* —the gain from keeping P_i closer to P_i^* —can be very small because a farmer's utility is insensitive to his relative price over a significant range. Thus, a small menu cost can produce a large x^* even if the resulting fluctuations in real output are highly inefficient. While excessive price flexibility is not likely to be an important problem, excessive rigidity may be.

III. Extensions

A. Heterogeneous Agents

In our basic model, multiple equilibria arise when each farmer chooses to adjust his price if and only if others do. The desire to make the same decision as others is crucial. A natural question is whether multiple equilibria are possible if heterogeneity leads some agents to adjust while others do not. Here we show that models with heterogeneity can possess multiple equilibria in the proportion of prices that adjust and, therefore, in the size of the real effects of a nominal shock. We focus on heterogeneity in the size of menu costs, which is the simplest case. Strategic complementarity is necessary for multiple equilibria; the sufficient condition depends on the distribution of the menu cost. Other sources of heterogeneity lead to similar results.

Assume that the menu cost, z , varies across farmers with cumulative distribution function $H(z)$. After a shock, farmers with z below some critical level adjust their prices, and the others do not. Let k be the proportion that adjust. We derive an equilibrium condition for k .

Let $P_A(x, k)$ be the price set by those who adjust and let $P(x, k)$ be the aggregate price level. Note that $P_A = P_i^* = P^\phi(1+x)^{1-\phi}$ and [approximating (6)] $P \cong kP_A + (1-k)$. These relations imply

$$(24) \quad P_A(x, k) \cong 1 + \frac{1-\phi}{1-\phi k} x.$$

By reasoning similar to that in Section II,

the gain from adjusting is

$$(25) \quad G(x, k) = V\left(\frac{1+x}{P(x, k)}, 1\right) - V\left(\frac{1+x}{P(x, k)}, \frac{1}{P(x, k)^\phi(1+x)^{1-\phi}}\right).$$

Using (24) and (25), one can show that

$$(26) \quad G(x, k) \cong -\frac{1}{2} \left(\frac{1-\phi}{1-\phi k} \right)^2 V_{22} x^2.$$

The crucial result is

$$(27) \quad \frac{\partial G(x, k)}{\partial k} > 0.$$

The gain from adjusting is increasing in the proportion of firms that adjust. This is a generalization of the earlier result that the gains are greater when all adjust than when none adjusts. Again, adjustment by others moves the price level in the same direction as the money supply, which increases the deviation of P_i^* from 1.

A farmer pays his menu cost if it is less than $G(x, k)$. Thus, the proportion who pay is $H(G(x, k))$, and an equilibrium k is one that satisfies $k = H(G(x, k))$. A necessary condition for multiple equilibria is $\partial H(G(x, k))/\partial k > 0$ over some range. Since $\partial H(G(x, k))/\partial k = dH/dG \cdot \partial G/\partial k$ and $H(\cdot)$ is increasing over some range, the condition reduces to (27), which holds because of strategic complementarity. The sufficient condition depends on the size of x and the shape of $H(\cdot)$; it is easy to find examples both of multiple equilibria and of unique equilibria.¹⁴

¹⁴Introducing heterogeneous real shocks leads to similar results. Suppose that the production function (3) is replaced by $Y_i = \theta_i L_i$, that θ_i varies across farmers, and that a shock to θ_i occurs at the same time as the monetary shock. Farmer i will choose to pay the menu cost if θ_i is above an upper cutoff or below a lower cutoff; both critical values depend on x and k . Again, one can show that multiple equilibria are possible and that strategic complementarity is a necessary condition.

Our results again parallel others in the coordination-failure literature. Diamond (1982), for example, introduces heterogeneity in the costs of production opportunities. Greater aggregate production raises an agent's incentive to produce by creating more trading partners. This strategic complementarity can produce multiple equilibria in the proportion of opportunities undertaken. As in our model, sufficient conditions depend on the distribution of costs.

B. Dynamics

So far we have studied a static model. In reality, price rigidity is a failure of prices to adjust quickly over time. Therefore, we now consider a dynamic version of our model. We focus on an example in which farmers choose between adjusting prices every period and adjusting every two periods; at the end, we briefly consider more general cases. There are two results. First, strategic complementarity in optimal prices produces multiple equilibria in the frequency of adjustment and hence in the dynamics of real output. Second, in contrast to other coordination-failure models, the economy converges to a unique long-run equilibrium.

Assume that the money stock follows a random walk; its innovations have mean zero and variance σ_m^2 . A farmer can adjust his price every period or every two periods, and he pays a menu cost z for each adjustment. When a farmer adjusts, he does so after observing the current money stock. If he adjusts every period, he always sets $P_i = P_i^*$. If farmers adjust every two periods, they all adjust in even periods; that is, price-setting is synchronized (Gary Fethke and Andrew Policano [1984] and Ball and Romer [1989b] show that this is the equilibrium timing when all shocks are aggregate). In this case, since M is a random walk, in even periods farmers set $P_i = P_i^* = M$; in odd periods, prices do not adjust to the most recent change in M .^{15,16}

¹⁵When prices are set for two periods, a farmer's optimal price in fact differs slightly from the first-period value of M , because this is not the optimal rigid price for the second period: as in the static model, certainty

We assume that each farmer chooses his frequency of adjustment taking others' frequency as given and solve for Nash equilibria. This exercise is a simple extension of the static case. A farmer compares the added cost of adjusting in odd periods to the expected gain from keeping $P_i = P_i^*$ in odd periods, which depends on others' frequency of adjustment. One can show that adjustment only in even periods is an equilibrium if

$$(28) \quad \frac{-(1-\phi)^2}{2} V_{22} \sigma_m^2 < z$$

and adjustment every period is an equilibrium if

$$(29) \quad -\frac{1}{2} V_{22} \sigma_m^2 > z$$

where we use approximations analogous to (14) and (17). Conditions (28) and (29) are the same as the conditions for nonadjustment and adjustment in the static model except that x^2 , the square of a given shock, is replaced by σ_m^2 , the *expected* square of the shock. The reason is that farmers decide how frequently to adjust before observing the realizations of shocks.

Since $0 < \phi < 1$, there are multiple equilibria in price adjustment for a range of z (or, for given z , for a range of σ_m^2). This multiplicity implies multiple equilibria in output dynamics. If prices adjust every period, monetary shocks are neutral, and output is constant. If prices adjust only in even periods, shocks in odd periods cause output movements that last until the next adjustment. In contrast to our static model, the

equivalence fails. Allowing farmers to choose prices different from the initial M introduces complications similar to the ones for the static model (see the Appendix).

¹⁶Rather than adjust every period, a farmer could guarantee $P_i = P_i^*$ by adjusting every two periods (or never) but indexing his price to the money stock. Thus, an alternative interpretation of the model is that a farmer chooses between setting a noncontingent price for two periods and setting an indexed price. Under this interpretation, z is an indexation cost.

output effect of an odd-period shock is strictly increasing in the size of the shock.¹⁷

Strategic complementarity is again the source of multiple equilibria. Intuitively, more frequent adjustment by others makes the price level respond more quickly to shocks and thus makes it more volatile. For $\phi > 0$, greater volatility in the price level implies greater volatility in a farmer's desired price, which increases his incentive to adjust frequently. For some parameter values, the incentive to adjust every period exceeds the added cost if and only if others adjust every period.

While we focus here on a simple example, the central results carry over to more general settings. An earlier version of this paper (Ball and Romer, 1988) considers a continuous-time model in which farmers can choose any frequency of adjustment. Strategic complementarity in desired prices can lead to multiple equilibria in the frequency. This implies multiple equilibria in the adjustment speed of the aggregate price level and hence in the path of output following a shock. Sufficient conditions for multiple equilibria depend on how steeply the costs of price adjustment increase with the frequency of adjustment.

Finally, our dynamic model makes clear a difference between coordination failure in price adjustment and coordination failures identified by previous authors. In previous models, there is no reason for an economy in a Pareto-dominated equilibrium to leave it. For example, if each agent in the Diamond model does not produce because others do not produce, this situation need not improve over time. In contrast, our model implies differences between the short-run and long-run behavior of the economy. Multiple equilibria in the frequency of price adjustment imply multiple equilibria in the

size and duration of the output effects of nominal shocks. However, there is a unique long-run response to a shock: prices eventually adjust fully, and the shock is neutral.¹⁸

IV. Conclusions and Implications

This paper shows that nominal price rigidity can arise from a failure of firms to coordinate price changes. Increases in price flexibility by different firms are strategic complements: greater flexibility of one firm's price raises the incentives for other firms to make their prices more flexible. Strategic complementarity can lead to multiple equilibria in the degree of nominal rigidity, and welfare may be much higher in the low-rigidity equilibria. Thus, the inefficient economic fluctuations resulting from nominal shocks might be greatly reduced if agents could "agree" to move to a superior equilibrium.

We conclude by discussing the empirical and policy implications of our results. One implication is that there can be considerable variation across economies in the degree of nominal rigidity and hence in the size of real fluctuations, without large variation in the underlying determinants of rigidity. Multiple equilibria imply that differences in rigidity can arise without any underlying differences. Even with unique equilibria, strategic complementarity implies that there is a "multiplier" (Cooper and John, 1988): small underlying differences can lead to large differences in rigidity. Nominal rigidity does in fact appear to vary considerably across countries; for example, Dennis Grubb et al. (1983) estimate that the adjustment of nominal wages to inflation is more than three times as fast in the average Western European country as in the United States. It could be a mistake to search for explanations of such differences based on the un-

¹⁷The result that money matters in odd but not even periods is unattractive, but it can be eliminated through realistic modifications of the model. For example, if idiosyncratic productivity or demand shocks arrive at different times for different farmers, then there can be an equilibrium with staggered adjustment: half of all prices change every period (Ball and Romer, 1989b). In this case, the effect of a shock does not depend on when it occurs, and the shock's real effects are strictly increasing in its size.

¹⁸If we modified the model so that the short-run response of the economy had permanent effects, through either capital accumulation or more exotic "hysteresis" mechanisms (Blanchard and Lawrence Summers, 1986), then the economy would no longer have a unique long-run equilibrium. Even in this case, however, there would be a unique long-run degree of price rigidity (full flexibility).

derlying natures of economies. Perhaps the apparent importance of unexplained "institutions" simply reflects the fact that different economies settle at different equilibria.

Another set of empirical implications arises if we ask why an economy arrives at one equilibrium rather than another. Current coordination-failure models generally do not address this subject, but a natural possibility is that the selection of an equilibrium is determined by history (Howard Naish, 1987; Lawrence Summers, 1987). For concreteness, consider the institutions governing wage-setting, such as the presence or absence of indexation. It appears natural to assume that these arrangements do not change with changing economic conditions as long as existing institutions continue to be among the set of equilibria. With this additional assumption, our model suggests that differences in wage-setting across similar economies can be explained by differences in past conditions; differences in past inflation variability, for example, might account for differences in the prevalence of indexation. It does not appear difficult to test for such an effect of history. One might, for example, regress a cross-country measure of the extent of indexation on current and historical variables.

Our results also add to the implications of menu-cost models for microeconomic data. One can test directly for strategic complementarity with data on the lengths of labor contracts or the frequency of changes in individual prices. The natural approach is to estimate the relation across countries or time periods between the frequency of an individual firm's wage or price adjustment and the average frequency in the economy. (Of course, there is an identification problem, since one firm's frequency could respond to the same unobservable variables as others' frequency; one would need to find instruments.) Such a test would be in the spirit of Aloysius Siow (1987), who uses microeconomic data to test for strategic complementarity in individuals' hours of work.

The result in previous papers that equilibrium rigidity can be excessive suggests a role for government regulation of price-set-

ting, such as restrictions on the lengths of labor contracts. This paper's results strengthen this policy implication in several ways. First, with multiple equilibria, policy can be less coercive. Instead of prohibiting certain contract provisions, the government could simply convene meetings of business and labor leaders to coordinate adjustment (as some European governments appear to do). Second, by moving the economy to a new equilibrium, temporary regulations can permanently change the degree of nominal rigidity. There is evidence of such effects: Stephen Cecchetti (1987) finds that the Nixon wage-price controls have permanently altered the provisions of U.S. labor contracts. Third, the multiplier arising from strategic complementarity magnifies the effects of policy. Regulation of union contracts would directly affect only a small fraction of wages in the United States. However, more flexible union wages would increase the incentives for wage and price flexibility throughout the economy and thus could have large effects on overall flexibility.

Finally, if coordination of wage and price adjustment proves difficult, an alternative is to substitute active monetary policy. In the *General Theory*, John Maynard Keynes (1936 pp. 266-8) argued that it is easier for the government to offset a fall in demand by increasing the money stock than for decentralized agents to reduce nominal wages in tandem. As Summers (1987) points out, governments adjust schedules through daylight saving time because it is difficult for decentralized agents to coordinate on a desirable equilibrium. Perhaps governments should be responsible for offsetting macroeconomic shocks for similar reasons.

APPENDIX

This appendix relaxes the assumption of our static model that all prices equal 1 before the monetary shock occurs. We assume instead that farmers choose initial prices optimally and show how this affects our results. The analysis draws heavily on Ball and Romer (1989a). As in that paper, we assume that the distribution of the monetary shock is symmetric around zero, single-peaked, and continuous.

The price that a farmer sets before observing the money supply depends on others' initial prices and on the value of the cutoff x^* . In symmetric equilibrium, each farmer's initial price is

$$(A1) \quad P_0(x^*) \cong 1 + \frac{\gamma}{2} \hat{\sigma}_M^2(x^*)$$

where $\hat{\sigma}_M^2(x^*)$ is the variance of M conditional on $1 - x^* < M < 1 + x^*$ (see Ball and Romer, 1989a). Given our use of second-order approximations, assuming initial prices equal to P_0 rather than 1 does not affect our results about equilibrium rigidity: the expressions for x_N and x_A in the text remain valid (Ball and Romer [1989a] shows this for x_N). We now show, however, that the socially optimal degree of rigidity changes slightly.

If initial prices are P_0 and prices are rigid, then $M/P = M/P_0$ and $P_i/P_i^* = P_0^{1-\phi}/M^{1-\phi}$. Thus, when initial prices are set optimally, a farmer's expected utility, (20), becomes

(A2)

$$E[U_i] = \{1 - [F(1+x^*) - F(1-x^*)]\} [V(1,1) - z] + \int_{m=1-x^*}^{1+x^*} V\left(\frac{M}{P_0(x^*)}, \frac{[P_0(x^*)]^{1-\phi}}{M^{1-\phi}}\right) f(M) dM.$$

The first-order condition for x_S is

$$(A3) \quad -[f(1-x_S) + f(1+x_S)][V(1,1) - z] + f(1-x_S)V\left(\frac{1-x_S}{P_0(x_S)}, \frac{[P_0(x_S)]^{1-\phi}}{(1-x_S)^{1-\phi}}\right) + f(1+x_S)V\left(\frac{1+x_S}{P_0(x_S)}, \frac{[P_0(x_S)]^{1-\phi}}{(1+x_S)^{1-\phi}}\right) + \left\{ \int_{M=1-x_S}^{1+x_S} \left[V_1\left(\frac{M}{P_0(x_S)}, \frac{[P_0(x_S)]^{1-\phi}}{M^{1-\phi}}\right) \times \left(\frac{-M}{[P_0(x_S)]^2}\right) + V_2\left(\frac{M}{P_0(x_S)}, \frac{[P_0(x_S)]^{1-\phi}}{M^{1-\phi}}\right) \times \left(\frac{(1-\phi)[P_0(x_S)]^{-\phi}}{M^{1-\phi}}\right) \right] f(M) dM \right\} \times P'_0(x_S) = 0$$

where $P'_0 \equiv (\partial P_0 / \partial x^*)$. Taking a second-order approximation and substituting (A1) for P_0 yields

$$(A4) \quad -[f(1-x_S) + f(1+x_S)] \times [V(1,1) - z] + f(1-x_S) \times \{V(1,1) + V_1[-x_S - (\gamma/2)\hat{\sigma}_M^2] + (1/2)V_{11}x_S^2 + (1/2)V_{22}(1-\phi)^2x_S^2\} + f(1+x_S)\{V(1,1) + V_1[x_S - (\gamma/2)\hat{\sigma}_M^2] + (1/2)V_{11}x_S^2 + (1/2)V_{22}(1-\phi)^2x_S^2\} - (\gamma/2)V_1[f(1+x_S) + f(1-x_S)] \times (x_S^2 - \hat{\sigma}_M^2) = 0.$$

Finally, using the fact that $f(1+x) = f(1-x)$, the solution for x_S is

$$(A5) \quad x_S \cong \sqrt{\frac{-2z}{V_{11} + (1-\phi)^2V_{22} - \gamma V_1}}$$

Substituting the derivatives of $V(\cdot)$ into (A5) establishes that $x_S < x_N$ and that x_S can be either greater or less than x_A . The possibility of $x_S < x_A$, which implies that all equilibria possess too much rigidity, is the main departure of these results from the ones in the text. The explanation is that P_0 , the price level under rigidity, is greater than 1, its level in the text. Thus, average output under rigidity is lower than in the text, which makes it more likely that reducing rigidity would increase welfare.

REFERENCES

- Akerlof, George and Yellen, Janet, "A Near-Rational Model of the Business Cycle, with Wage and Price Inertia," *Quarterly*

- Journal of Economics*, Supplement 1985, 100 (5), 823-38.
- Azariadis, Costas and Cooper, Russell, "Nominal Wage Rigidity as a Rational Expectations Equilibrium," *American Economic Review*, May 1985 (*Papers and Proceedings*), 75, 31-5.
- Ball, Laurence, "Externalities from Contract Length," *American Economic Review*, September 1987, 77, 615-29.
- _____ and Romer, David, "Sticky Prices As Coordination Failure," mimeo, Princeton University, 1988.
- _____ and _____, (1989a) "Are Prices Too Sticky?" *Quarterly Journal of Economics*, August 1989, 104, 507-24.
- _____ and _____, (1989b) "The Equilibrium and Optimal Timing of Price Changes," *Review of Economic Studies*, April 1989, 56, 179-98.
- Blanchard, Olivier, "Why Does Money Affect Output? A Survey," NBER (Cambridge, MA) Working Paper No. 2285, June 1987; *Handbook of Monetary Economics*, forthcoming.
- _____ and Kiyotaki, Nobuhiro, "Monopolistic Competition and the Effects of Aggregate Demand," *American Economic Review*, September 1987, 77, 647-66.
- _____ and Summers, Lawrence, "Hysteresis and the European Unemployment Problem," in Stanley Fischer, ed., *NBER Macroeconomics Annual*, Cambridge, MA: MIT Press, 1986, pp. 15-77.
- Bryant, John, "A Simple Rational Expectations Keynes-Type Model," *Quarterly Journal of Economics*, August 1983, 98, 525-9.
- Cecchetti, Stephen, "Indexation and Incomes Policy: A Study of Wage Adjustment in Unionized Manufacturing," *Journal of Labor Economics*, July 1987, 5, 391-412.
- Cooper, Russell and John, Andrew, "Coordinating Coordination Failures in Keynesian Models," *Quarterly Journal of Economics*, August 1988, 103, 441-63.
- Diamond, Peter, "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, October 1982, 90, 881-94.
- Farmer, Roger E. A. and Woodford, Michael, "Self-Fulfilling Prophecies and the Business Cycle," CARESS Working Paper No. 84-12, University of Pennsylvania, April 1984.
- Fethke, Gary and Policano, Andrew, "Wage Contingencies, the Pattern of Negotiation, and Aggregate Implications of Alternative Contract Structures," *Journal of Monetary Economics*, September 1984, 14, 151-71.
- Grubb, Dennis, Jackman, Richard and Layard, Richard, "Wage Rigidity and Unemployment in OECD Countries," *European Economic Review*, March/April 1983, 21, 11-39.
- Keynes, John Maynard, *The General Theory of Employment, Interest, and Money*, London: Macmillan, 1936.
- Kiyotaki, Nobuhiro, "Multiple Expectational Equilibria under Monopolistic Competition," *Quarterly Journal of Economics*, November 1988, 103, 695-713.
- Mankiw, N. Gregory, "Small Menu Costs and Large Business Cycles," *Quarterly Journal of Economics*, May 1985, 100, 529-37.
- McCallum, Bennett, "On 'Real' and 'Sticky-Price' Theories of the Business Cycle," *Journal of Money, Credit, and Banking*, November 1986, 18, 397-414.
- Naish, Howard, "The Costs of Rational Pricing Behavior, and the Non-Neutrality of Money," mimeo, Pitzer College, 1987.
- Rotemberg, Julio J., "The New Keynesian Microfoundations," in Stanley Fischer, ed., *NBER Macroeconomics Annual*, Cambridge, MA: MIT Press, 1987, pp. 69-104.
- Siow, Aloysius, "Coordinating Hours of Work," mimeo, Columbia University, 1987.
- Summers, Lawrence, "Should Keynesian Economics Dispense with the Phillips Curve?" in Rod Cross, ed., *Unemployment, Hysteresis, and the Natural Rate Hypothesis*, Oxford: Blackwell, 1987, pp. 11-25.
- Woodford, Michael, "Indeterminacy of Equilibrium in the Overlapping Generations Model: A Survey," mimeo, Columbia University, May 1984.

When Excessive Consumption Is Rational

By RICHARD E. ROMANO*

If average cost is everywhere above market demand, it is usually argued that the nondiscriminating firm will shut down, although the first-best outcome may dictate production. In this setting, it is shown that there is often a Nash equilibrium in consumption that will keep the firm producing. Selfish consumers engage in excessive (beyond demand) consumption to keep the firm in business and to protect their surpluses. This is shown to be true in a simple model with perfect information and also in a more realistic model in which consumers are uncertain about the firm's costs. (JEL L10, D80)

"You [resident of Lake Wobegon] need a toaster, you buy it at Co-op Hardware even though you can get a deluxe model with all the toaster attachments for less money at K-Mart in St. Cloud."

[Garrison Keillor, 1985 pp. 95-6]

Have you "small towners" ever found yourselves in the following type of situation? A new restaurant opens that has an unusual menu but one that well suits your preferences. You suspect the local demand to be slack and fear for the restaurant's survival. You contemplate more frequent patronage than you would if you were certain that the place would not fail.

This paper shows that such behavior can be rational. It can be optimal to consume in excess of that indicated by one's demand curve, derived under standard conditions, in particular when it is believed that one can purchase all that is desired over the market period. This "excessive consumption" has nothing to do with altruism. It can be a part of a Nash equilibrium in consumption with

selfish consumers. Though Keillor attributes the behavior described in the above quotation to social-mindedness, I will offer a competing explanation.

The logic of the result is the following. There is an incentive to consume beyond demand to protect one's surplus when nonexcessive consumption may be insufficient to keep a firm producing. Keeping the firm in business is a public good, so that there is a countervailing incentive to free ride on others' excessive consumption. If consumers know the firm's costs, then they know exactly the level of consumption that is necessary to keep the firm producing; and the incentive to free ride is resolved in a Nash consumption equilibrium. If consumers are uncertain about the firm's costs but fear that the firm may go out of business, free riding may not be perfectly resolved in Nash equilibrium, although there is still excessive consumption. Since excessive consumption is voluntary, it obviously will result in improved efficiency.

I. The Certain Case

A. An Example

A simple example illustrates an excessive-consumption equilibrium and highlights the issues. The example is illustrated in Figure 1. Consider a market with two identical consumers who have demands $d_i = \frac{1}{2}(a - p)$, $i = 1, 2$, so that market demand is given by $d = a - p$. There is a single firm with cost

*Department of Economics, University of Florida, Gainesville, FL 32611. My thanks to Sandy Berg, Roger Blair, Thomas Cooper, Jonathan Hamilton, Shmuel Nitzan, Steve Slutsky, John Tschirhart, Ed Zabel, and two anonymous referees for helpful comments and suggestions, though I retain responsibility for any errors. This research was supported by a grant from the Public Policy Research Center at the University of Florida to which I am indebted.



FIGURE 2. TIMING OF THE TWO-STAGE, ONE-PERIOD GAME PLAYED BY CONSUMERS AND THE FIRM

expressed consumption levels are inadequate; and, likewise, the firm anticipates the resulting Nash consumption levels when it sets price. I analyze a kind of "subscription equilibrium," where no consumption takes place unless the subscribed consumption levels are sufficiently large. An alternative would be to analyze a repeated game.¹ The present rendition is simple in that it permits analysis of a one-period game.

The Results.—The first proposition regards equilibrium in the consumption stage of the game. Assumption 1 is adopted in the three propositions of this subsection.

PROPOSITION 1: (a) $q^* = (q_1^*, q_2^*, \dots, q_n^*)$ is a Nash equilibrium with positive consumption if and only if (i) $q_i^* \geq d_i(p)$, (ii) $CS_i \geq 0$, and (iii) $\Pi(p, \Sigma q_i^*) = 0$. (b) In any Nash equilibrium with positive consumption, at least one consumer engages in excessive consumption [i.e., $q_i^* > d_i(p)$ for some i].

PROOF:

(a) Consider first the necessity of (i)–(iii). If $q_i^* < d_i(p)$, then q_i^* is not a best response. Likewise, if $CS_i < 0$, then q_i^* is not a best response. Regarding the necessity of (iii), if $\Pi < 0$, then positive consumption will not result. If $\Pi > 0$, then $q_i^* > d_i(p)$ for some i by Assumption 1, and that consumer could reduce consumption without causing $\Pi < 0$ and thereby could increase CS_i . Hence, $\Pi = 0$ is also necessary. Sufficiency of (i)–(iii) follows from the fact that no consumer gains by deviating from q_i^* . Increasing q_i lowers CS_i by (i). Decreasing q_i causes the firm not to produce by (iii), and the consumer would sacrifice CS_i .

(b) $q_i^* > d_i(p)$ for some i or, otherwise, (iii) and Assumption 1 are inconsistent.

An interesting characteristic of an excessive-consumption equilibrium is that, even though there are externalities in consumption (all consumers would lose if one reduces consumption from equilibrium), free riding is controlled. The situation provides a natural and harsh punishment to such deviations in the form of exit and lost consumer surplus.² Consumers are willing to engage in excessive consumption to protect the surplus obtained on the units below ordinary demand. If price is such that there exists an excessive-consumption equilibrium, there is an infinite number of such equilibria (excepting the boundary case in which $CS_i = 0$ for all i in equilibrium). To see this, suppose that q^* is an excessive-consumption equilibrium with elements (q_i^*, q_j^*) , $CS_i > 0$, and $q_j^* > d_j$. Then, the modified vector of consumption levels with elements $(q_i^* + \varepsilon; q_j^* - \varepsilon)$ is another excessive-consumption equilibrium. Note that the aggregate consumption level is constant across such equilibria.

In addition, there are, in general, trivial equilibria with zero consumption. Such equilibria are Pareto inferior to equilibria with positive consumption, and they will be ignored in what follows.³ This leads to the

²These Nash equilibria in consumption are similar to those obtained in Bengt Holmstrom's (1982 p. 327) theorem 2. Likewise, they resemble the equilibria that arise in the literature on the voluntary provision of discrete public goods (Thomas R. Palfrey and Howard Rosenthal, 1984; Mark Bagnoli and Barton L. Lipman, 1989). The problem with certain costs provides a natural example of the voluntary provision of a discrete public good.

³One might also proceed under the assumption that the firm attaches some probability to the occurrence of a "zero-consumption equilibrium" when pricing. This would not change the results in this section (see footnote 7, below, regarding the next section), since profits are zero in such an equilibrium and also in an equilibrium with positive consumption.

¹A longer version of this paper discusses this and other extensions. Contact the author for a copy.

determination of price, which is the topic of Proposition 2.

PROPOSITION 2: *The firm is indifferent over its choice of price.*

PROOF:

This follows immediately from the fact that profits equal zero in any excessive-consumption equilibrium.

There will generally be a range of prices that support excessive-consumption equilibria, and the firm is indifferent among these. Since profits equal zero in equilibrium, the firm is also indifferent about entering. This bothersome result will not be true in the more realistic case examined in the next section.

An excessive-consumption equilibrium is obviously Pareto superior to the firm not entering. It is also obvious that there must be positive net surplus at the first-best consumption/output vector for there to exist an excessive-consumption equilibrium. This is also sufficient for the existence of equilibrium if consumers have identical preferences, but it is not otherwise. Proposition 3 formalizes these results and presents the necessary and sufficient condition under which an excessive-consumption equilibrium exists. Two definitions are useful. Let $q_i^f \equiv d_i^{-1}(c)$, so that $\mathbf{q}^f = (q_1^f, q_2^f, \dots, q_n^f)$ is the first-best consumption vector. Define $\bar{q}_i(p)$ implicitly in $CS_i(p, \bar{q}_i) = 0$: $\bar{q}_i(p)$ equals the maximum quantity a consumer would ever purchase in an excessive-consumption equilibrium.

PROPOSITION 3: (a) *The condition*

$$(1) \quad \sum CS_i(p, q_i^f) + \Pi(p, \sum q_i^f) \geq 0$$

is necessary for there to exist an excessive-consumption equilibrium. (a') Condition 1 is also sufficient if consumers are identical but is not otherwise. (b) The existence of a p such that

$$(2) \quad (p - c) \sum \bar{q}_i(p) \geq F$$

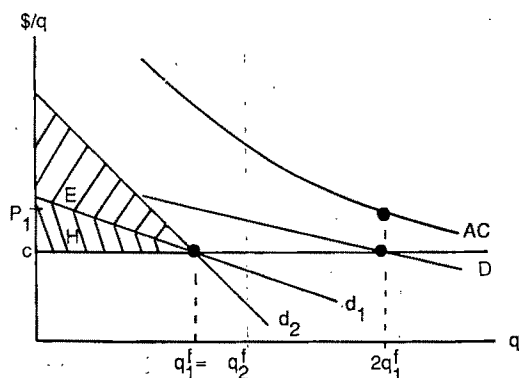


FIGURE 3. A MARKET WITH TWO CONSUMERS WITH DEMANDS d_1 AND d_2

is necessary and sufficient for there to exist an excessive-consumption equilibrium.

PROOF:

(a) If $\sum CS_i(p, q_i^f) + \Pi(p, \sum q_i^f) < 0$, then there exists no vector \mathbf{q} with some positive consumption levels such that the expression is nonnegative. Hence, it would be impossible to satisfy conditions (ii) and (iii) of Proposition 1 with any positive consumption levels.

(a') Let $\Pi(p^f, \sum q_i^f) = 0$ define p^f . For the case of identical consumers, since $CS_i(p^f, q_i^f) = CS_j(p^f, q_j^f)$ for all i and j , the augmented vector (p^f, \mathbf{q}^f) is easily seen to be an excessive-consumption equilibrium.

A counterexample may be used to show that (1) is not sufficient in the presence of nonidentical consumers. Refer to Figure 3 which depicts a market with two consumers with demands d_1 and d_2 . D is the market demand. Assume that the indicated areas E and H satisfy $E + 2H = F$. Thus, (1) is satisfied with equality, but no excessive-consumption equilibrium exists, as is now argued. Let p_1 satisfy $\Pi(p_1, q_1^f + q_2^f) = 0$, and note that $CS_1(p_1, q_1^f) + CS_2(p_1, q_2^f) + \Pi(p_1, q_1^f + q_2^f) = CS_1(p_1, q_1^f) + CS_2(p_1, q_2^f) = 0$. It is clear from the figure that $CS_2(p_1, q_2^f) > CS_1(p_1, q_1^f)$, and therefore, the latter is negative. This precludes an excessive-consumption equilibrium with the first-best values. Since (1) vanishes, there can be no equilibrium at other than the first-best val-

ues, since the "available surplus" would be negative.

(b) To see necessity of existence of a p such that (2) is satisfied, suppose none exists. Then, the requirement of an excessive-consumption equilibrium that $\Pi = 0$ necessitates $q_i > \bar{q}_i$ for at least one consumer; but such a choice could not be a best response, since this implies $CS_i < 0$.

Sufficiency of the condition is shown by construction. For any p such that (2) is satisfied, there will exist a vector q with $q_i \leq \bar{q}_i(p)$ such that (2) holds with equality. Moreover, there exists such a vector with $q_i \geq d_i(p)$, or else Assumption 1 is contradicted. Any such vector is easily seen to be an excessive-consumption equilibrium.

Proposition 3 indicates that a potential for surplus is enough to guarantee that some surplus can be realized in an excessive-consumption equilibrium, provided consumers have identical preferences. This is not sufficient without identical preferences, because of the following. The model imposes the constraint that every consumer must pay the same per unit of consumption. This implies that every consumer's *average* valuation must not be less than average cost at the first best, if it is to be sustainable as a Nash equilibrium.⁴ Identical preferences are required to ensure this.

⁴Several of my colleagues have suggested that I relate this analysis to the all-or-none demand curve. An individual's all-or-none demand curve shows the maximum price he would be willing to pay to obtain each quantity. There are two ways one might construct a *market* all-or-none demand curve; both assume zero income effects. One construction horizontally sums the individual's all-or-none demand curves. For this construction, there will exist an excessive-consumption equilibrium if and only if the market all-or-none demand curve crosses the average cost curve. Alternatively, using the ordinary market demand curve, one could calculate the total value placed on each market quantity and then relate each quantity to the corresponding average value. This will be the same as the first construction if consumers have identical preferences, but it will otherwise generally be beyond the first construction, because it does not require each consumer to have the same average valuation. It is necessary but not sufficient that the second construction crosses the average cost curve for there to exist an excessive-consumption equilibrium.

With or without identical preferences, the proposition should not be understood to imply that the first best is the compelling outcome. It is worth noting that there can exist an excessive-consumption equilibrium when the first-best consumption vector is not sustainable as a Nash equilibrium.⁵ Moreover, note that there is only one price that could sustain the first-best consumption vector as an equilibrium and yet the firm is indifferent over price.

An excessive-consumption equilibrium has the razor's-edge property that a marginal reduction in consumption by any consumer will cause the firm to exit. This stretches one's imagination, since it is difficult to believe that any consumer in the real world will have enough information to imagine himself on such a precipice. The natural question to ask is whether excessive consumption will persist in a model with a more realistic information structure. The answer is shown to be "yes" in the next section, where F will be unknown to consumers when the Nash consumption game is played. The razor's-edge property of the certainty model is also related to some troubling outcomes of the model. One such outcome is that the model fails to predict both a unique price and consumption pattern because of the two types of multiple equilibria that arise. Also, one finds that the firm is indifferent to entering. The extension will go a long way toward resolving these problems.

II. The Case of Uncertainty About Fixed Costs

To capture consumer uncertainty, I will assume that consumers do not observe the fixed cost in this section. The main results are twofold. Excessive consumption will continue to characterize Nash equilibrium; and under fairly general conditions, equilibrium will be unique. After describing the modifications to the model, the results are described, and an example is presented.

⁵Contact the author for an example of this.

A. Modifications to the Model

The timing is the same as in Figure 2, except that only the firm observes F . The fixed cost is drawn from a continuous probability distribution $G(F)$ with support $[a, b] \in [0, \infty)$. The associated density function g is strictly positive and differentiable in the interior of its support. Further, the following is assumed.

ASSUMPTION 2: *The ratio $g(x)/G(x)$ is nonincreasing in x .*

Assumption 2 is useful in establishing the existence and uniqueness properties of equilibrium. It is satisfied by the set of Polya frequency distributions of order two (see Richard E. Barlow and Frank Proschan, 1965), which is a large set of unimodal densities including (with appropriate restrictions on parameters) the uniform, exponential, beta, gamma, Weibull, normal, and truncated normal distributions. Though consumers will not observe the realization of F , G is assumed to be common knowledge.

Note that $G[(p - c)\sum q_i]$ is the probability that the realization of fixed cost is no greater than revenue minus variable costs, which then equals the probability that the firm produces. The analogous assumption to Assumption 1 here is the following.

ASSUMPTION 1':

$$G[(p_m - c)\sum d_i(p_m)] < 1$$

where p_m is again the usual monopoly price.

Hence, I assume that there is a positive probability that the firm would shut down with monopoly pricing and consumption determined by ordinary demands. A final assumption is as follows.

ASSUMPTION 3:

$$G[(p - c)\sum \bar{q}_i(p)] > 0 \text{ for some } p$$

where, recall, $CS_i(p, \bar{q}_i) = 0$ defines $\bar{q}_i(p)$.

This is analogous to (2) in part (b) of Proposition 3.⁶ Assumptions 1', 2, and 3 are maintained throughout this section. Everything else, including the notation, is the same as in the certainty case.

B. Analysis and Results

Consider first the consumption stage of the game. Consumers will maximize their expected consumer surplus, taking into account their influence on the probability that the firm chooses to produce. The i th consumer chooses $q_i \geq 0$ to maximize the function

$$(3) \quad E[CS_i] = \left(\int_0^{q_i} d_i^{-1}(x) dx - pq_i \right) \times G[(p - c)\sum q_j].$$

Three lemmas are useful.

LEMMA 1: *$E[CS_i]$ is strictly quasi-concave in q_i for all $\sum_{j \neq i} q_j$ and $G > 0$.*

PROOF:

See the Appendix.

LEMMA 2: *For p such that Assumption 3 is satisfied, there exists a Nash equilibrium (in pure strategies) with $G > 0$.*

PROOF:

See the Appendix.

Lemma 2 indicates that there will be a nontrivial equilibrium with a positive probability of production for some prices. There may also exist equilibria with zero probability of resulting production (e.g., the zero consumption vector may be an equilibrium). As in the certainty case, these equilibria will be Pareto inferior to any equilibrium with a positive probability that the firm will produce. I ignore these equilibria in what fol-

⁶Given the other assumptions, Assumption 3 will be necessary and sufficient for the existence of an excessive-consumption equilibrium, though the paper proceeds in a somewhat different order here.

lows.⁷ Note that Lemma 2 implies that there exists an equilibrium with excessive consumption if, for all prices, there is zero probability that the firm would produce with consumption determined by ordinary demand (i.e., if $G[(p_m - c)\sum d_i(p_m)] = 0$). I will show that the weaker Assumption 1' is sufficient for the existence of an equilibrium with excessive consumption.

LEMMA 3: *A Nash consumption equilibrium will never have $(p - c)\sum q_i > b$ (recall that b is the upper bound of the support of F).*

PROOF:

By Assumption 1', an equilibrium with $(p - c)\sum q_i > b$ would require that $q_i > d_i$ for some i . But such a choice could not be a best response. The consumer could always marginally decrease consumption, maintain $G = 1$, and thereby increase utility.

The last obstacle to describing Nash consumption equilibria is to note that the payoff functions in (3) are nondifferentiable in q_i at $(p - c)\sum q_j = b$ if $g(b) > 0$. This leads to the possibility of either an "interior equilibrium" or a "corner equilibrium," which are respectively described by the systems

$$(4a) \quad (d_i^{-1}(q_i^*) - p)G[(p - c)\sum q_j^*] \\ + (p - c)CS_i(p, q_i^*) \\ \times g[(p - c)\sum q_j^*] = 0 \\ i = 1, \dots, n$$

⁷As in the certainty case (see footnote 3), one could also proceed by assuming that the firm attaches some probability to the result of a trivial equilibrium, where such equilibria exist. If it is assumed that this probability is independent of price, which may be reasonable if there exist trivial equilibria at all relevant prices ($p > c$), then the results of this section are unaffected. Alternatively, one might somehow link this probability to some measure of the relative number of such equilibria, in which case the results require some modification. It seems sensible to attach a higher probability to Pareto superior equilibria, and I have chosen to assign a probability of 1 to this set.

$$(4b) \quad a < (p - c)\sum q_j^* < b$$

or

$$(5a)$$

$$d_i^{-1}(q_i^*) - p + (p - c)CS_i(p, q_i^*)g(b) \geq 0 \\ i = 1, \dots, n$$

$$(5b) \quad (p - c)\sum q_j^* = b.$$

In both cases, without loss of generality, I ignore consumers who demand zero of the good at the current price [I assume $d_i(p) > 0$ for all i].

The first proposition of this section describes further the nature of equilibrium and the conditions under which the two types arise and, most importantly, shows the existence of excessive consumption. Its statement is facilitated by defining $\hat{q}_i(p)$ as the value that satisfies

$$(6) \quad d_i^{-1}(\hat{q}_i) - p + (p - c)g(b) \\ \times CS_i(p, \hat{q}_i) = 0.$$

This equals the maximum quantity a consumer would be willing to consume in a Nash equilibrium when it is certain that the firm will produce. Note that $\hat{q}_i \geq d_i$.

PROPOSITION 4: *Assume that price is such that Assumption 3 is satisfied. (a) If $(p - c)\sum \hat{q}_i(p) < b$, then there is a unique Nash equilibrium with a positive probability that the firm will produce. It has $0 < G < 1$, and every consumer engages in excessive consumption. (b) If $(p - c)\sum \hat{q}_i(p) > b$, then there are multiple Nash equilibria with positive probability that the firm produces. Each has $G = 1$ and at least one consumer who engages in excessive consumption.*

PROOF:

See the Appendix.

The corner equilibria in the uncertainty model are similar to the excessive-consump-

tion equilibria that arise in the certainty model. Where one arises, there is an infinite number, each of which entails excessive consumption by a subset of consumers. Proposition 4 indicates that there will either be multiple corner equilibria or a unique interior equilibrium.⁸ The interior equilibrium may, however, be a more "likely" outcome. Before discussing further the properties of an interior equilibrium, the next proposition makes a case for their prevalence.

PROPOSITION 5: *Assume that price is such that Assumption 3 is satisfied. Any of the following conditions implies that an interior equilibrium arises:*

- (a) $\sum CS_i(c, q_i^f) < b$, where $q_i^f \equiv d_i(c)$ is i 's first-best consumption level;
- (b) $g(b) = 0$;
- (c) in a finite economy ($n < \infty$), the range of F exceeds some finite value (that is a function of parameters); in a finite economy, the variance of F exceeds some finite value.

PROOF:

(a) This is shown by demonstrating that there cannot be an equilibrium with $G = 1$. If there were, then from (5b), $(p - c)\sum q_i^* = b$. That $\sum CS_i(c, q_i^f) < b$ implies $\sum CS_i(p, q_i) + (p - c)\sum q_i < b$ for any p and vector q and, in particular, any equilibrium vector q^* . Hence, $\sum CS_i(p, q_i^*) < b - (p - c)\sum q_i^* = 0$. This implies $CS_i(p, q_i^*) < 0$ for some i ; and since $E[CS_i] = CS_i G = CS_i$ here, it implies that some consumer has negative expected consumer surplus in equilibrium. Of course, this cannot happen.

(b) If $g(b) = 0$, then (6) implies $\hat{q}_i = d_i$, and thus Assumption 1' implies $(p - c)\sum \hat{q}_i < b$. Hence, the result follows from part (a) of Proposition 4.

(c) Since $a \geq 0$, b eventually increases without bound as the range increases. Hence, the result follows from part (a) of

this proposition. Now consider the variance (σ^2):

$$\begin{aligned}\sigma^2 &\equiv \int_a^b (F - E[F])^2 dG(F) \\ &< (b - a)^2 \int_a^b dG(F) \\ &= b^2 + a^2 - 2ab < 2b^2\end{aligned}$$

where the last inequality uses the nonnegativity of a . The result then follows by the same argument as for the range.⁹

Part (a) of Proposition 5 asserts that, if there is some probability that the fixed cost will exceed the maximum net surplus that could be obtained from producing the good, then (nontrivial) Nash consumption equilibrium will be unique. The same is true by part (b) of Proposition 5 if uncertainty is such that g vanishes at its upper bound. Note that this can be true with arbitrarily little uncertainty; for example, consider an exponential distribution of F (but with a bounded away from zero for realism) with arbitrarily small variance. Part (c) of Proposition 5 shows that the same is true in a finite economy if there is enough uncertainty present.

The incentive to protect one's surplus continues to influence consumer choice in the uncertainty model. This incentive is present as long as the probability that the firm will produce is less than 1. This incentive confronts every consumer in an interior equilibrium, and every consumer engages in excessive consumption. Since increasing the probability that the firm produces is a public good, this result may seem surprising in light of the extant literature on provision of public goods in Nash contribution games (see Theodore Bergstrom et al. [1986] and the references therein). A result of this lit-

⁸If $(p - c)\sum \hat{q}_i(p) = b$, then $q_i^* = \hat{q}_i$ is the unique equilibrium.

⁹An alternative proof shows that $g(b)$ converges to zero as the variance grows (for a fixed mean) and then appeals to part (b) of Proposition 5. This is notable, since it does not require that $\sum CS_i(c, q_i^f)$ is finite, thus applying to a market with an infinite number of consumers.

erature is that free-riding is very likely (see James Andreoni, 1988; Timothy L. Fries et al., 1991). The relevant difference is that here the marginal cost to consumers of increasing their consumption beyond demand begins at zero. Hence, every consumer engages in excessive consumption. In contrast, the marginal cost of making a monetary contribution to a public good is 1, and only those individuals for whom the marginal benefit exceeds 1 will contribute positively.

In a corner equilibrium, it is possible that the excessive consumption by a subset of consumers resolves the problem of probabilistic provision for all consumers. Here, not all consumers will necessarily engage in excessive consumption in equilibrium. Motivated by Proposition 5, I consider further only the case in which an interior equilibrium arises.¹⁰ Hence, assume any one of the conditions of Proposition 5 prevails in all that follows.

There are two externalities of consumption in an interior equilibrium. First, one consumer's increased consumption benefits other consumers by increasing the probability that the firm produces. Second, the firm obviously benefits. These externalities prevent an interior equilibrium from being socially efficient. The next proposition formalizes this point. I adopt expected welfare

$$\begin{aligned}
 (7) \quad E[W] &= \Sigma E[CS_i] + E[\Pi] \\
 &= G[(p-c) \Sigma q_i] \Sigma CS_i(p, q_i) \\
 &\quad + \int_a^{(p-c)} \Sigma q_i [(p-c) \Sigma q_i - F] dG(F)
 \end{aligned}$$

as the social-welfare measure. In addition, I will use the fact that equilibrium price must exceed c , which is shown below.

PROPOSITION 6: *An interior equilibrium is never socially efficient.*

¹⁰A longer version of this paper (see footnote 1) examines further the properties of corner equilibria. Also, see Shmuel Nitzan and Romano (1990), who examine Nash contribution games for discrete public goods with uncertain costs.

PROOF:

In an interior equilibrium, $\partial E[CS_i]/\partial q_i = 0$. It is straightforward to check that $\partial E[CS_j]/\partial q_i > 0$ for all $j \neq i$; and, using that $p > c$, $\partial E[\Pi]/\partial q_i > 0$. Then, $E[W]$ can be increased by increasing q_i .

The force of the externalities worsens as the economy grows. The consequence of this is seen by examining the following replication. Let n increase and keep the fixed cost *per customer* constant. Correspondingly, reinterpret F as the fixed cost per customer and replace $G(F)$ with $G(F/n)$ in the above analysis. It is straightforward to confirm that the equilibrium probability of production approaches that without any excessive consumption as n increases. Hence, near free riding prevails in large economies, and the empirical relevance of this analysis is limited to cases with sufficiently small numbers.

The firm chooses price anticipating the outcome of the consumption stage. To analyze this, let $N(p) \equiv \Sigma q_i^*(p)$, which is unique under the assumptions. The firm's price choice maximizes its expected profits, which equal $\int_a^{(p-c)N(p)} [(p-c)N(p) - F] dG(F)$. This problem is equivalent to

$$(8) \quad \max_p (p-c)N(p)$$

or, also, maximization of the probability of ultimate production. The final proposition regards this maximization.

PROPOSITION 7: (a) *The firm will have positive expected profits in equilibrium.* (b) *The optimal price may be uniquely determined.*

PROOF:

(a) The firm produces only for realizations of F in which it earns nonnegative profit, and so the result follows from the fact that $G > 0$ at the optimum. (b) See the example in the following subsection.

In contrast to the certain case, the firm strictly prefers entering here. Price is not always uniquely determined, because $N(p)$

need not be well behaved.¹¹ It has proved futile to search for meaningful conditions that guarantee uniqueness of price. It is easy, however, to provide an example in which $N(p)$ is well behaved and price is unique.

C. An Example

Let cost be given by $C = q + F$, where F has a uniform distribution with support $[0, b]$. There are two identical consumers with demands $d_i = A - p$, $i = 1, 2$, where $A > 1$. It is easy to check that Assumptions 2 and 3 are satisfied. Since $p_m = (A + 1)/2$, $G[(p_m - c)\sum d_i(p_m)] = (A - 1)^2/2b$, and thus Assumption 1' requires that $b > (A - 1)^2/2$. The i th consumer's expected consumer surplus is given by

$$\begin{aligned} E[CS_i] &= \left[\int_0^{q_i} (A - x) dx - pq_i \right] \\ &\quad \times G[(p - 1)(q_i + q_j)] \\ &= (p - 1)(q_i + q_j) \\ &\quad \times [(A - p)q_i - q_i^2/2] / b \\ &\quad i = 1, 2, i \neq j \end{aligned}$$

where it is assumed that $G < 1$ (which conforms to an interior equilibrium). Maintaining the latter assumption for the moment, the Nash equilibrium values satisfy

$$\begin{aligned} (9) \quad &2(A - p)q_i - 3q_i^2/2 \\ &+ q_j(A - p - q_i) = 0 \\ &i = 1, 2, i \neq j. \end{aligned}$$

The unique solution to (9) is $q_i^* = q_j^* = 6(A - p)/5$, which implies $N(p) = 12(A - p)/5$. Then, the firm maximizes $(p - 1)N(p)$, with solution $p^* = (A + 1)/2$. All this assumes that $G[(p - c)N(p)]$ is less than 1 for all p . This requires $b > 3(A - 1)^2/5$.

III. Summary and Concluding Remarks

The economic setting in which average cost lies beyond demand is traditionally regarded as one in which the good will not be produced, although this may be suboptimal. I have shown that there is often a Nash equilibrium in consumption in which the good is produced in this setting. Individuals consume beyond their demands to keep the firm in business and to protect their surpluses. With perfect information, there are multiple equilibria; and the model fails to predict uniquely price and the consumption pattern. By introducing consumer uncertainty about cost, the multiple-equilibrium problem often disappears. Moreover, all that is required to generate excessive consumption is consumers' perception of some chance that the firm will exit without such behavior.

In the more realistic uncertain case, excessive consumption declines as the economy is replicated. This limits the empirical relevance of the model. The effect of excessive consumption is likely to be significant only in small local economies. Some examples of local monopolies resulting from large fixed costs relative to market size are ethnic and fast-food restaurants, speciality food and beverage stores, fresh-fish markets, professional stage theaters, and taxi and airline services. The presence of local monopoly does not itself provide evidence of excessive consumption. It is casual empiricism that motivated this study. I have, for example, observed the somewhat reluctant purchase of season tickets to the local theater despite the availability of tickets to every play. An empirical test for excessive consumption could be constructed as follows. The model with uncertainty about the fixed cost also predicts that an individual's consumption decreases in the number of consumers (without replication of fixed cost) once the market is small enough that there is a positive probability of exit.¹² The model then

¹¹A longer version of this paper (see footnote 1) contains an example in which $N(p)$ has an upward-sloping range.

¹²This can be shown using (4a) and Assumption 2. A proof is in a longer version of this paper (see footnote 1).

predicts a negative coefficient on sufficiently small market sizes in a regression that explains an individual's consumption.

Enough said, as I am on my way (again) to a mediocre Chinese restaurant, but the only one in town that serves a good Peking Duck.

APPENDIX

PROOF OF LEMMA 1:

Taking derivatives yields

$$(A1) \quad \frac{\partial E[CS_i]}{\partial q_i} = (d_i^{-1} - p)G + (p - c)gCS_i$$

$$= [d_i^{-1} - p + (p - c)CS_i g / G]G$$

and

$$(A2) \quad \frac{\partial^2 E[CS_i]}{\partial q_i^2} = [d_i^{-1} - p + (p - c)CS_i g / G]g + d_i^{-1} + (p - c) \left[(d_i^{-1} - p)g / G + CS_i \frac{\partial(g/G)}{\partial q_i} \right]$$

Strict quasi-concavity is implied if $E[CS_i]$ is strictly concave whenever $\partial E[CS_i] / \partial q_i$ vanishes, since then there exists no interior minimum. From (A1) and (A2), if $\partial E[CS_i] / \partial q_i = 0$, then the sign of $\partial^2 E[CS_i] / \partial q_i^2$ equals the sign of what follows the first term in (A2). Recalling that g/G is nonincreasing, the result then follows if $CS_i \geq 0$ and $d_i^{-1} - p \leq 0$. Now, $d_i^{-1} - p > 0$ implies $CS_i > 0$. If these hold, one can see by inspection that (A1) cannot vanish. The other possibility would have $d_i^{-1} - p \leq 0$ and $CS_i < 0$. Again (A1) could not vanish. Hence, the result follows.

PROOF OF LEMMA 2:

I show that there exists an equilibrium with $G > 0$ in an artificially restricted game and then show it to have properties implying that it is also an equilibrium in the game of interest. Let the Cartesian product of strategy sets in the restricted game be given by $S = \{q: (p - c)\sum q_i \geq a \text{ and, for all } i, 0 \leq$

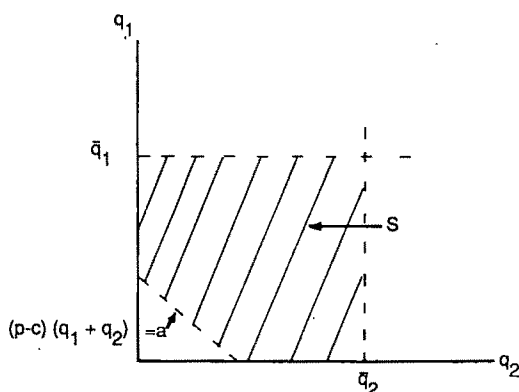


FIGURE 4. A RESTRICTED GAME WITH TWO CONSUMERS

$q_i \leq \bar{q}_i$). Figure 4 illustrates a case of two consumers. The first restriction on the vector q is the "artificial one," (i.e., the game of interest drops this requirement). Since p satisfies Assumption 3 in the text, S has an interior. Also, S is closed and convex. Then, since the i th payoff function is continuous and quasi-concave in q_i (Lemma 1), the choice correspondence satisfies the hypotheses of Kakatani's fixed-point theorem; and hence there will exist an equilibrium in the restricted game. See James W. Friedman (1986 pp. 36-9) for details.

Now I show that no equilibrium in the restricted game can occur on the "south-westerly bound" of S where $(p - c)\sum q_i = a$ and G vanishes. Let $\bar{q} = (\bar{q}_1, \bar{q}_2, \dots, \bar{q}_n)$ denote a choice on this bound. Since G would vanish, $E[CS_i] = 0$ for all i . But, for any $q_i \in (\bar{q}_i, \bar{q}_i)$, $E[CS_i] > 0$, implying that \bar{q}_i cannot be a best response. Hence, equilibrium must occur elsewhere in S and, in particular, have $G > 0$. Any equilibrium in the restricted game must then also be an equilibrium in the unrestricted game. The unrestricted game adds only strategy space where $E[CS_i] = 0$, which will never contain a better response than a best response in S . Hence, there always exists an equilibrium in the unrestricted game with $G > 0$.

PROOF OF PROPOSITION 4:

(a) First I show that there exists an equilibrium with $0 < G < 1$. Since it is known

from Lemma 2 that there exists an equilibrium with $G > 0$, it suffices to show that there is no equilibrium with $G = 1$. Suppose that there is such an equilibrium. Then, since $(p - c)\Sigma \hat{q}_i < b$, $q_i^* > \hat{q}_i$ for some i . However, from Lemma 1 and equation (6), this implies that q_i^* cannot be a best response [(5a) would not be satisfied for i]. Hence, any equilibrium must have $0 < G < 1$ and satisfy (4a) and (4b).

To show the remaining results, rewrite (4a) as

$$(A3) \quad \frac{d_i^{-1}(q_i^*) - p}{(p - c)CS_i(p, q_i^*)} = - \frac{g[(p - c)Q^*]}{G[(p - c)Q^*]}$$

where $Q^* \equiv \Sigma q_i^*$. Since the right-hand side of (A3) is negative, it must be that $d_i^{-1} - p < 0$. This establishes that $q_i^* > d_i$ (i.e., every consumer engages in excessive consumption). Equation (A3) is of the form $\phi_i(q_i) = \psi(Q)$, where

$$\phi_i' = \frac{d_i^{-1}(p - c)CS_i - (d_i^{-1} - p)^2(p - c)}{(p - c)^2 CS_i^2} < 0$$

and $\psi' \geq 0$ since g/G is nonincreasing (recall Assumption 2). Therefore, $q_i = H_i(Q)$, where $H_i(Q) \equiv \phi_i^{-1}(\psi(Q))$. Then,

$$(A4) \quad Q^* = \Sigma H_i(Q^*).$$

Since $\phi_i' < 0$ and $\psi' \geq 0$, $H_i' \leq 0$ for each i , implying that (A4) can have no more than one solution. Thus, equilibrium is unique.

(b) I show first that there exists an equilibrium with $G = 1$ by using Lemma 2 and demonstrating that no equilibrium with $0 < G < 1$ exists. Since $(p - c)\Sigma \hat{q}_i(p) > b$, the latter such equilibrium would require that $q_i^* < \hat{q}_i$ for some i . From Lemma 1 and equation (6), then (4a) would not be satisfied (i.e., q_i^* could not be a best response).

To see the remaining results, note that using Assumption 1' yields

$$(p - c) \Sigma d_i(p) < b < (p - c) \Sigma \hat{q}_i(p).$$

It is clear then that there exist multiple vectors with the properties $q_i \in [d_i, \hat{q}_i]$ and $\Sigma q_i = b$, each of which must have $q_i > d_i$ for some i . Using (5a), (5b), (6), and Lemma 1, it is straightforward to check that each of these is an equilibrium vector, and it is clear that there can be no others. That $q_i > d_i$ for some i implies that at least one consumer engages in excessive consumption.

REFERENCES

- Andreoni, James, "Privately Provided Public Goods in a Large Economy: The Limits of Altruism," *Journal of Public Economics*, February 1988, 35, 57-73.
- Bagnoli, Mark and Lipman, Barton L., "Provision of Public Goods: Fully Implementing the Core through Private Contributions," *Review of Economic Studies*, October 1989, 56, 583-602.
- Barlow, Richard E. and Proschan, Frank, *Mathematical Theory of Reliability*, New York: Wiley, 1965.
- Bergstrom, Theodore, Blume, Lawrence and Varian, Hal, "On the Private Provision of Public Goods," *Journal of Public Economics*, February 1986, 29, 25-40.
- Friedman, James W., *Game Theory and Applications to Economics*, New York: Oxford University Press, 1986.
- Fries, Timothy L., Golding, Edward and Romano, Richard E., "Private Provision of Public Goods and the Failure of the Neutrality Property in Large Finite Economies," *International Economic Review*, February 1991, 32, 147-58.
- Holmstrom, Bengt, "Moral Hazard in Teams," *Bell Journal of Economics*, Autumn 1982, 13, 324-40.
- Keillor, Garrison, *Lake Wobegon Days*, New York: Viking Penguin, 1985.
- Nitzan, Shmuel and Romano, Richard E., "Private Provision of a Discrete Public Good with Uncertain Costs," *Journal of Public Economics*, August 1990, 42, 357-70.
- Palfrey, Thomas R. and Rosenthal, Howard, "Participation and the Provision of Discrete Public Goods: A Strategic Analysis," *Journal of Public Economics*, July 1984, 24, 171-93.

Capital Formation and Productivity Convergence Over the Long Term

By EDWARD N. WOLFF*

Catch-up in total factor productivity (TFP) among the "group of seven" was evident between 1870 and 1979, though much slower before 1938 than after 1950. Capital:labor ratios also converged over the long period, though the process was much stronger after 1960. TFP catch-up is found to be positively associated with capital:labor growth and strongest when capital intensity is growing most rapidly. The United States overtook the United Kingdom in technological leadership in 1900 when its capital:labor growth was more than three times higher. The steady deterioration in the United Kingdom's relative TFP since 1900 and the United States' since 1950 are both associated with low rates of capital formation. (JEL O57, O30, J24, O40)

Recent studies have documented a convergence both in average labor productivity levels and in per capita income over the last century or so and particularly since the end of World War II among industrialized economies (see e.g., Moses Abramovitz [1979, 1986], Angus Maddison [1982, 1987], Gottfried Bombach [1985], and William Baumol [1986] for labor productivity statistics; Baumol and Wolff [1988] for GDP per capita; and Steve Dowrick and Duc-Tho Nguyen [1989] for total factor productivity). Abramovitz's (1986) and Baumol's (1986) results, in particular, highlight these trends. They found an almost perfect inverse relation between labor productivity levels in 1870 and the rate of labor productivity growth between 1870 and 1979 among 16 Organization for Economic Cooperation and Development (OECD) countries. In addition, the coefficient of variation in productivity levels, defined as the ratio of the standard deviation to mean productivity, fell from 0.48 in 1870 to 0.16 in 1979.

Abramovitz (1986) also investigated sub-periods and found that labor productivity convergence was much slower in the period before World War II than after. Indeed, even in the postwar period, there is evidence from Abramovitz and from Baumol and Wolff (1988) that productivity convergence has slowed down during the 1970's, though this is disputed by Dowrick and Nguyen, who find parameter stability in their catch-up model between pre- and post-1973 periods when controlling for factor-intensity growth. Abramovitz also found that there were significant changes in leadership and the rank order of countries over time. Results of Bradford De Long (1988) show very little evidence of productivity convergence over the last century when the sample is no longer restricted to OECD countries. However, Baumol and Wolff, using the Robert Summers and Alan Heston (1988) sample, which covers countries at all levels of development, found convergence in real GDP per capita among the top third or so over the 1950-1981 period, though it was weaker than among OECD countries alone.

Explanations of the productivity catch-up almost all involve the so-called "advantages of backwardness," by which it is meant that much of the catch-up can be explained by the diffusion of technical knowledge from the leading economies to the more backward ones (see e.g., Alexander

*Department of Economics, New York University, New York, NY 10003. I thank the National Science Foundation, the Exxon Foundation, and the C. V. Starr Center for Applied Economics for support of the research. I also thank William Baumol, Moses Abramovitz, and four anonymous referees for their valuable suggestions.

Gerschenkron, 1952; Simon Kuznets, 1973). Indeed, the further an economy is from the technological frontier, the greater the rate of technical advance possible from such borrowing. However, being backward does not itself guarantee that a nation will catch up. Other factors must be present, such as strong investment (see Abramovitz [1979 p. 2] for a discussion of potential advances in productivity from capital accumulation), an educated work force, a suitable product mix, and developed trading relations with advanced countries. This paper investigates the role of capital formation in the process of productivity catch-up.

This study considers three hypotheses (which are not mutually exclusive) to account for the observed convergence in labor productivity levels among the advanced nations. The first is the "catch-up" hypothesis, which states that countries that lag furthest behind the leading countries in terms of technology level should exhibit the most rapid rate of growth in technology. This would also imply convergence in total factor productivity (TFP) levels, defined as the ratio of output to a weighed sum of labor and capital inputs, among nations. The second hypothesis is that the convergence in labor productivity levels has been due to narrowing of differences in factor intensities (capital:labor ratios) among industrialized countries.

The third hypothesis is that there are positive interactions between capital accumulation and technological advance. This deserves some comment. There are several avenues through which capital formation and total factor productivity growth may be associated. First, it is likely that substantial capital accumulation is necessary to put new inventions into practice and to effect their widespread employment. This association is often referred to as the "embodiment effect," since it implies that at least some technological innovation is embodied in capital. It is also consistent with the "vintage effect," which states that new capital is more productive than old capital per (constant) dollar of expenditure. If the capital stock data do not correct for vintage effects, then a positive correlation should be observed

between the rate of technological gain and the change in the growth rate of capital.

A second avenue is that the introduction of new capital may lead to better organization, management, and the like. This may be true even if no new technology is incorporated in the capital equipment. A third avenue is through learning-by-doing (see Kenneth Arrow, 1962). Thus, technological advance should be correlated with the accumulation of capital stock. Fourth, potential technological advance may stimulate capital formation, because the opportunity to modernize equipment promises a high rate of return to investment. A fifth avenue is through the so-called Verdoorn or Kaldor effect, whereby investment growth may lead to a growth in demand and thereby to the maintenance of a generally favorable economic climate for investment. Such positive feedbacks may act cumulatively. These last four arguments do not lead to a specific functional relation between TFP growth and the rate of capital or capital:labor growth but do suggest a positive correlation between the two sets of variables.

In my analysis, it is not possible to distinguish among these various effects, and I will refer to them collectively as interaction effects or complementarities between capital accumulation and technological advance. Moreover, as is apparent, it is not possible to attribute causation one way or the other, since the influence between TFP growth and capital formation runs in both directions. However, it is possible to test for these interaction effects, and results will be reported below.¹

The empirical analysis is limited to the "group of seven"—Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States—over the 1870–1979 period, because of the availability of consistent data on total capital stock for this period provided in Maddison (1982). This sample is by no means representative and is

¹Also, see Richard Nelson (1964), Kuznets (1973), Abramovitz and Paul David (1973), Bombach (1985), and Robert Solow (1988) for related discussions.

subject to many of the same criticisms made by De Long (1988) of the OECD samples.

Support is found for the three hypotheses. First, TFP levels converged among the group of seven over the 1870–1979 period. However, the pattern is far from uniform, with convergence much stronger after World War II than before, as Abramovitz (1986) found for labor productivity. Second, aggregate capital:labor ratios showed convergence over the long period, though the process was much stronger after 1960; before World War II, it is evident only when the United States, which surged ahead of the other countries between 1900 and 1938, is excluded from the sample. Technological advance and capital formation played about equal roles in labor productivity growth.

Third, the data show a positive correlation of 0.79 between the rate of TFP growth and that of the capital:labor ratio over the 1880–1979 period. Results, based on regression analysis and a vintage model, are somewhat mixed but generally support the existence of an interaction effect between technological advance and capital accumulation. The effect was strongest during the postwar period, when both capital:labor growth and the speed of technological catch-up were greatest. Overall, convergence in labor productivity levels is found to be a consequence of all three effects.

The next part of the paper provides evidence of convergence in productivity levels, and Section II presents statistics on growth in capital and changes in capital intensity over time. Section III considers the relation between capital formation and the rate of technological progress. Concluding remarks are made in the last section.

I. Productivity Catch-Up

The TFP level for country h is defined as the ratio of total output (Y^h) to a weighted average of labor input (L^h) and capital input (K^h)

$$(1) \quad TFP^h = Y^h / [\alpha^h L^h + (1 - \alpha^h) K^h]$$

where the labor input is measured by hours of work, the capital stock is measured by

gross nonresidential fixed plant and equipment, and α^h is the wage share in country h .² Total factor productivity growth is based on the Divisia measure, ρ , defined as

$$(2) \quad \rho^h = \hat{Y}^h - \alpha^h \hat{L}^h - (1 - \alpha^h) \hat{K}^h$$

where a superscript “hat” ($\hat{}$) indicates the relative rate of change (see Frank Gollup and Dale Jorgensen [1980] for a discussion of the Divisia index). The Tornqvist approximation based on average period shares is employed.³

The choice of the proper factor shares is debatable. Under the assumption that technology is the same across countries and that factor prices are equalized, as in the Heckscher-Ohlin model (see e.g., Edward Leamer, 1984) factor shares should be equal among countries, so that international average factor shares provide the best approximation. On the other hand, if technology differs among countries, then country-specific wage shares should be used. The results point to the latter choice. However, the only available data for the full 1870–1979 period are employee compensation (EC) and national income (NI) for the United Kingdom and the United States, so that factor shares are based on the average ratio of EC to NI in the two countries.⁴

²John Kendrick and R. Sato (1963) demonstrate that this index can be derived as a special case of the CES production function.

³Two other indexes were employed. First, the translog index of TFP level [see equation (4) in Section III] was used as an alternative to equation (1). Second, the time derivative of equation (1) was used as an alternative to the Divisia index. Results are similar to those based on (1) and (2) and are not reported here.

⁴For the postwar period, data availability is much greater, and therefore several alternative measures of factor shares were constructed, including the ratio of EC to GDP, the inclusion of a labor portion of entrepreneurial (self-employment income) in the wage share, and country-specific factor shares. In addition, net capital stock estimates were also available for all countries except Italy. Furthermore, since, productivity movements are sensitive to business-cycle fluctuations, the TFP index was also adjusted for capacity utilization, as

$$(1') \quad TFPU^h = Y^h / [\alpha^h L^h + (1 - \alpha^h) u^h K^h]$$

TABLE 1—TOTAL FACTOR PRODUCTIVITY (TFP) LEVELS, 1870–1979

Country	TFP (Index numbers, United States = 1.00 in 1950) ^a										
	1870	1880	1890	1900	1913	1929	1938	1950	1960	1970	1979
Canada	—	—	—	—	—	0.50	0.49	0.86	1.00	1.17	1.23
France	—	—	—	—	—	—	—	0.54	0.78	1.12	1.31
Germany	0.16	0.18	0.22	0.27	0.31	0.38	0.46	0.43	0.76	1.01	1.12
Italy	—	0.17	0.18	0.20	0.26	0.34	0.42	0.45	0.65	1.04	1.21
Japan	—	0.08	0.09	0.11	0.14	0.24	0.32	0.21	0.36	0.83	1.01
United Kingdom	0.28	0.32	0.36	0.40	0.44	0.53	0.57	0.72	0.85	1.05	1.15
United States	0.25	0.30	0.36	0.41	0.49	0.65	0.65	1.00	1.13	1.28	1.38
Five-country statistics (Germany, Italy, Japan, United Kingdom, and United States):											
CV ^b	—	0.43	0.43	0.42	0.38	0.34	0.24	0.48	0.33	0.14	0.10
Max/min ^c	—	4.10	3.89	3.66	3.47	2.74	2.04	4.70	3.11	1.55	1.37
Average/U.S. ^d	—	0.62	0.59	0.60	0.59	0.57	0.69	0.45	0.58	0.77	0.81
Seven-country statistics:											
CV ^b	—	—	—	—	—	—	—	0.42	0.29	0.12	0.09
Max/min ^c	—	—	—	—	—	—	—	4.70	3.11	1.55	1.37
Average/U.S. ^d	—	—	—	—	—	—	—	0.54	0.65	0.81	0.85
Average annual rate of productivity growth											
Country	1880–1938			1950–1979			1880–1979				
	TFP (percentage)	Labor (percentage)	Ratio	TFP (percentage)	Labor (percentage)	Ratio	TFP (percentage)	Labor (percentage)	Ratio		
Canada	—	—	—	1.44	2.58	0.56	—	—	—		
France	—	—	—	3.04	4.64	0.66	—	—	—		
Germany	1.20	1.86	0.65	3.33	5.52	0.60	1.56	2.66	0.59		
Italy	0.58	1.80	0.32	3.09	4.99	0.62	1.36	2.59	0.52		
Japan	1.32	2.52	0.52	4.17	6.92	0.60	1.62	2.98	0.54		
United Kingdom	0.75	1.16	0.65	1.48	2.85	0.52	1.09	1.78	0.61		
United States	0.77	1.88	0.41	1.36	2.30	0.59	1.37	2.26	0.60		
Mean:	0.92	1.84	0.51	2.56	4.26	0.59	1.40	2.45	0.57		

^aTFP levels are computed according to equation (1). Output is measured by GDP, labor by hours worked, and capital by gross nonresidential fixed plant and equipment (net for Germany). Factor shares are based on the average ratio of employee compensation to national income for the United Kingdom and the United States over the 1870–1979 period.

^bCoefficient of variation, defined as the ratio of the standard deviation to the mean.

^cRatio of the maximum to the minimum productivity level.

^dRatio of unweighted average productivity level of all countries except the United States to U.S. productivity level.

Table 1 shows computations of TFP levels for the period from 1870 to 1979.⁵ It should be noted that the sample of countries diminishes as one goes further back in time because of data availability. The United

Kingdom was the early leader in total factor productivity. The United States caught up to the United Kingdom by 1890 and led

where u is the capacity utilization rate, based on the utilization index for the manufacturing sector. Results did not materially differ from those reported here and are not shown.

⁵The primary data source on output, gross capital stock, and hours worked is Maddison (1982), in which

problems of comparability of measures across countries are discussed. Estimates of GDP and man-hours for the 19th century are based on partial data. Many of the estimates are performed through backward interpolation of average growth rates. This has the effect of smoothing out the series and biasing the results toward convergence. Random errors of measurement at period end points will also likely bias the results in favor of convergence (see Abramovitz [1986] for a discussion of

thereafter. The United Kingdom remained in second place until 1950, when it was overtaken by Canada; by 1979, it had fallen to fifth (out of seven countries). Japan was last throughout the period, though its TFP relative to the United States increased from one-fourth in 1880 to three-fourths in 1979.

this point). Input measures do not adequately capture differences in natural resources, particularly land. This is particularly problematic for the early years (1870–1913), when the economies of these countries had a large agricultural sector. Because of the land:labor-ratio advantage of the United States and the declining share of agriculture in GDP over time of all seven countries, the technological gains of the United States relative to other countries will be understated by the TFP estimates. Differences in human capital are not reflected in the data. Differences in service-life assumptions, retirement patterns, depreciation schedules, and capital prices among countries will lead to inconsistencies in capital stock estimates among countries. Maddison attempted to standardize the estimates by benchmarking each national capital stock series to specially constructed 1976 estimates for each country, using international capital prices and the same assumptions with regard to service life and retirement patterns. Remaining errors in measurement will have the effect of overstating the variance of capital:labor ratios across countries and understating the variation of TFP relative to labor productivity. Also, see Abramovitz (1986) for further discussion of weaknesses in the long-term data.

Other data sources are as follows. Estimates of net capital stock are derived from Raymond Goldsmith (1985). The 1950 figures are from table 18 of Goldsmith; estimates for other years are based on individual country tables and geometric interpolation. Data on utilization rates are from the OECD's *Main Economic Indicators, 1960–1979* and David Coe and Gerald Holtham (1983). Data on wage shares are computed from the following sources: (i) data for 1950–1979 are from the United Nations' *Yearbook of National Accounts Statistics*, selected years, except for the 1950–1960 period in Italy; (ii) data for 1937–1950 and for 1950–1960 in Italy are from the International Labour Organization's *Yearbook of Labor Statistics*, various years; (iii) for Japan, data for 1920–1937 are from Kasushi Ohkawa and Henry Rosovsky (1973 pp. 316–7); (iv) for the United Kingdom, data for 1870–1938 are from Phyllis Deane and W. A. Cole (1964 p. 247); (v) for the United States, data for 1870–1938 are from D. Gale Johnson (1954); (vi) data on entrepreneurial income are from the United Nations' *Yearbook of National Accounts Statistics*, various years. The apportionment of entrepreneurial income into a wage and profit component is based on Johnson (1954), who estimated a 65-percent share for labor compensation and a 35-percent share for return on capital.

According to three indexes (the ratio of maximum to minimum TFP levels, the coefficient of variation, and the average TFP level of the other countries relative to the United States), there was only moderate convergence between 1880 and 1929 (particularly between 1880 and 1913). This is similar to labor productivity movements among the five countries.⁶ The Depression years did bring some convergence in TFP levels, followed by a sharp increase in dispersion between 1938 and 1950. This was partly a consequence of the deleterious effect of World War II on German and Japanese productivity, which declined in absolute terms, but mainly due to a tremendous increase in U.S. productivity. Another indicator of catch-up is a negative correlation of TFP growth rates with initial TFP levels. These coefficients show the same pattern: -0.20 for 1880–1913, -0.33 for 1880–1929, -0.64 for 1880–1938, and -0.83 for 1880–1979.

The postwar period (1950–1979 here) provides a relatively good case study of the catch-up process, because it is the longest stretch of time unbroken by a major war or a depression. Over this period, the coefficient of variation fell by more than two-thirds, the ratio of maximum to minimum TFP level declined by about two-thirds, and average TFP relative to the United States rose from 0.54 to 0.85. Moreover, the correlation of TFP growth rates with initial TFP levels (in 1950) was -0.96 .⁷

Results are also shown for average annual rates of both TFP and labor productivity growth. TFP growth averaged 1.4 percent per year over the 1880–1979 period,

⁶However, the results differ from Abramovitz's (1986) finding for labor productivity trends among 16 OECD countries. In particular, the coefficient of variation in labor productivity levels decreased from 0.51 in 1870 to 0.48 in 1880, 0.33 in 1913, and 0.29 in 1929. Thus, slow declines were observed for 1870–1880 and 1913–1929.

⁷Correlations were equally strong using the alternative measures of TFP introduced in footnote 4. Computations of correlation coefficients that exclude Japan were quite similar in magnitude. Correlations between TFP growth and the natural logarithm of initial TFP were even stronger.

and labor productivity growth averaged 2.5 percent per year. Japan led in both, while the United Kingdom was last. The ratio of TFP to labor productivity growth, a rough measure of the contribution of technical change to labor productivity growth, averaged 0.57, with the remaining portion due to capital deepening. There was relatively little variation among the five countries in the sample.

Over the 1880–1938 period, there were greater differences among countries. A comparison of the United Kingdom and the United States is revealing. Both experienced the same TFP growth. However, labor productivity growth was substantially lower in the United Kingdom, because of a much smaller growth in capital intensity. In fact, only a third of the United Kingdom's labor productivity growth was attributable to capital-deepening, in comparison to 60 percent for the United States. Over the postwar period, the ratio of TFP to labor productivity growth averaged 0.59, compared to 0.51 during 1880–1938, and there was less variation among countries.⁸

In sum, the period before World War II was one of moderate growth in both TFP and labor productivity and moderate catch-up in TFP levels. In contrast, the postwar period was characterized by strong growth in productivity and rapid convergence in productivity levels.

II. Capital Intensity

There are two apparently conflicting sets of results with regard to convergence in aggregate capital:labor ratios. According to the three summary measures shown in Table 2, there was slightly increasing disparity in capital:labor ratios before World War II and rapid convergence after 1960.⁹ How-

ever, correlations between initial capital:labor ratio and its rate of growth show catch-up in capital intensity: -0.28 for 1880–1913, -0.59 for 1913–1938, 0.24 for 1938–1950 and -0.91 for 1950–1979. For the full 1880–1979 period, the correlation was -0.97 . Thus, except for the period covering World War II, countries with lower initial levels of capital intensity experienced faster growth in their capital:labor ratio. The discrepancy is due to the very rapid growth in U.S. capital intensity between 1890 and 1938, by which time its capital:labor ratio was three times the average of the others. Indeed, when the United States is eliminated from the sample, the coefficient of variation in capital intensity shows a decline from 0.68 in 1880 to 0.37 in 1938.

The United Kingdom was the most capital-intensive in 1870 and 1880. The United States led in capital:labor growth during the 1880–1913 period and by 1890 was the most capital-intensive country, a position it held through 1970 (by 1979, Germany had become the most capital-intensive). U.S. capital-intensity growth was more than three times greater than the United Kingdom's between 1880 and 1913, which explains the emergence of the United States as the leader in labor productivity.

Three important relations become apparent from the data. First, there is a direct correspondence by period between the degree of capital-intensity catch-up and TFP convergence: strongest in 1950–1979, second strongest in 1913–1938, weakest in 1880–1913, and divergent in 1938–1950. Second, there is also a direct correspondence by period between TFP convergence and the average growth in capital intensity: highest in 1950–1979 (average capital:labor growth of 4.4 percent per year), next in 1913–1938 (2.2 percent), third in 1880–1913 (1.9 percent), and last in 1938–1950 (1.2 percent).¹⁰ Third, countries with higher cap-

⁸Maddison (1987) computed a higher average ratio of 0.74 over the 1950–1973 period for six countries: France, Germany, Japan, the Netherlands, the United Kingdom, and the United States. The difference is mainly due to Maddison's higher labor shares.

⁹Results for the postwar period are quite similar for the ratio of net capital stock to hours of work.

¹⁰A regression of the coefficient of variation of TFP levels on the average capital:labor growth rate for each of ten periods yields a highly significant negative coefficient on capital:labor growth.

TABLE 2—CAPITAL:LABOR RATIOS, 1870–1979

Country	Ratio of gross capital to hours (index numbers, standardized so that GDP per hour = 1.00 for United States in 1950) ^a										
	1870	1880	1890	1900	1913	1929	1938	1950	1960	1970	1979
Canada	—	—	—	—	—	1.34	1.46	1.89	2.82	3.60	4.45
France	—	—	—	—	—	—	—	1.22	1.58	2.60	4.08
Germany	0.30	0.38	0.47	0.59	0.74	0.87	0.89	0.99	1.46	2.87	5.12
Italy	—	0.14	0.18	0.21	0.30	0.46	0.68	0.75	0.98	1.95	3.23
Japan	—	0.06	0.07	0.08	0.12	0.23	0.28	0.32	0.40	1.18	2.55
United Kingdom	0.48	0.55	0.56	0.62	0.72	0.92	0.93	1.11	1.35	2.17	3.12
United States	0.41	0.52	0.60	0.87	1.23	1.75	2.14	2.41	3.15	4.06	4.89
Five-country statistics (Germany, Italy, Japan, United Kingdom, and United States):											
Mean	—	0.33	0.38	0.47	0.62	0.85	0.99	1.12	1.47	2.45	3.78
CV ^b	—	0.59	0.57	0.61	0.62	0.62	0.63	0.63	0.63	0.40	0.27
Average/U.S. ^c	—	0.54	0.54	0.43	0.38	0.36	0.32	0.33	0.33	0.50	0.72
Seven-country statistics											
Mean	—	—	—	—	—	—	—	1.24	1.68	2.63	3.92
CV ^b	—	—	—	—	—	—	—	0.52	0.54	0.35	0.23
Average/U.S. ^c	—	—	—	—	—	—	—	0.43	0.45	0.59	0.77
Average annual growth rates (percentages)											
Country	1880–1913		1913–1938		1938–1950		1950–1979		1880–1979		
Canada	—		—		2.17		2.96		—		
France	—		—		—		4.17		—		
Germany	2.07		0.69		0.92		5.67		2.64		
Italy	2.26		3.27		0.75		5.05		3.15		
Japan	1.92		3.57		1.08		7.12		3.76		
United Kingdom	0.83		1.02		1.47		3.56		1.75		
United States	2.61		2.23		0.98		2.44		2.27		
Mean:	1.94		2.16		1.23		4.42		1.50		
SD:	0.60		1.16		0.47		1.52		0.38		
CV: ^b	0.31		0.54		0.39		0.34		0.26		

^aThe labor input is measured by hours worked and capital by gross nonresidential fixed plant and equipment (net for Germany). Calculations of the mean, standard deviation, and coefficient of variation are based on countries in the sample with the relevant data.

^bCoefficient of variation, defined as the ratio of the standard deviation to the mean.

^cRatio of unweighted average for all countries except the United States to the U.S. value.

ital:labor growth generally had higher TFP growth. The rank order is identical for the postwar period: Japan (annual capital:labor growth of 7.1 percent), Germany (5.7 percent), Italy (5.1 percent), France (4.2 percent), United Kingdom (3.6 percent), Canada (3.0 percent), and the United States (2.4 percent). For the whole 1880–1979 period, Japan was first in capital:labor growth, followed by Italy, Germany, the United States, and the United Kingdom. Except for a reversal between Italy and Germany, the rank order was identical to that of TFP

growth. These empirical findings suggest the existence of interaction effects between capital growth and technology growth.

III. Interactions Between Capital Accumulation and Productivity Growth

I use a standard growth-accounting framework to assess the extent to which convergence in technology levels is associated with capital accumulation through interaction effects between the two. Formally, assume that for each country h there

is a Cobb-Douglas value-added production function:

$$(3) \ln Y^h = \zeta^h + \alpha \ln L^h + (1 - \alpha) \ln K^h.$$

The parameter ζ^h is country-specific and indicates country h 's technology level. The output elasticity of labor, α , is assumed to be the same across countries. If factors are paid their marginal products, then the output elasticity is equal to labor's distributional share. This study will use the cross-country (unweighted) average of labor's share as the estimate of α .

Next, define the translog index of TFP level:

$$(4) \ln TFP^h = \ln Y^h - \alpha \ln L^h - (1 - \alpha) \ln K^h$$

which is consistent with the Divisia index of TFP growth. Comparison of equation (4) with equation (3) reveals that this measure of TFP level is implicitly based on a Cobb-Douglas form for the production function. Moreover, let the United States be the benchmark country, and define the following:

π^h : ratio of country h 's labor productivity to U.S. labor productivity;

τ^h : ratio of country h 's technology level to U.S. technology level;

κ^h : ratio of country h 's capital:labor ratio to U.S. capital:labor ratio.

Equations (3) and (4) then imply that

$$(5) \ln \pi^h = \ln \tau^h + (1 - \alpha) \ln \kappa^h.$$

Differentiating this with respect to time yields

$$(6) \hat{\pi}^h = \hat{\tau}^h + (1 - \alpha) \hat{\kappa}^h.$$

Capital formation may be expected to exert two distinct effects on labor productivity. First, by raising the capital:labor ratio, it will increase labor productivity even if there is no advance in technology in use [equation (6)]. Second, through interactions with technology advance, accumulation may be associated with gains in productivity over and

above capital deepening. Three approaches for testing the interaction effect are considered here.

A. Correlation Between TFP and Capital:Labor Growth

The first and most direct test is to determine whether there is a positive correlation between $\hat{\tau}$ and $\hat{\kappa}$. Though this approach is not consistent with a strict vintage model, it probably captures the general set of interactions between the two variables, as discussed in the introduction. The correlation coefficient was 0.08 for 1880–1913, 0.37 for 1913–1938, 0.55 for 1938–1950, 0.95 for 1950–1979 and 0.79 for the whole 1880–1979 period.¹¹ These results are generally consistent with the hypothesis that high rates of technical advance are associated with high rates of capital formation. However, they indicate that the relation was considerably stronger after World War II than during the prewar years.

B. Regression Analysis

A second approach uses a regression framework. Two basic specifications are employed:

$$(7a) \hat{\tau}_t^h = b_0 + b_1 \tau_t^h + b_2 \hat{\kappa}_t^h + \varepsilon_t^h$$

$$(7b) \hat{\tau}_t^h = b_0 + b_1 \tau_t^h + b_2 \Delta \hat{\kappa}_t^h + \varepsilon_t^h$$

where τ_t^h is country h 's (translog) TFP relative to the United States at the start of each period, $\Delta \hat{\kappa}_t^h \equiv \hat{\kappa}_t^h - \hat{\kappa}_{t-1}^h$, and ε is a stochastic error term. In some specifications, country dummy variables (except the United Kingdom) are included to control for country-specific effects, such as the degree of trade openness, culture, and government policy. In some, period dummy variables are also included to allow TFP growth to vary by period (e.g., in response to unevenness in the flow of new technology or inventions).

¹¹Other periods were used in the analysis, including 1880–1900, 1900–1913, 1900–1929, 1913–1929, 1929–1938, and 1929–1950, with similar results.

TABLE 3—REGRESSIONS OF RELATIVE PRODUCTIVITY GROWTH ($\hat{\tau}$) ON RELATIVE PRODUCTIVITY LEVEL AND CAPITAL:LABOR GROWTH AND INTENSITY, 1880–1979

Independent variable	Regression						
	1	2	3	4	5	6	7
Constant	−0.011* (2.55)	−0.011* (2.55)	−0.006 (1.42)	−0.008 (1.47)	−0.001 (0.21)	−0.011 (1.64)	−0.002 (0.34)
τ	−0.042** (3.94)	−0.039** (3.60)	−0.042** (3.45)	−0.080** (5.72)	−0.085** (5.07)	−0.078** (5.00)	−0.073** (4.04)
$\hat{\kappa}$		0.130 (1.15)		0.189* (2.09)		0.395** (3.47)	
$\Delta\hat{\kappa}$			0.008 (0.65)		0.003 (0.79)		0.024* (2.17)
Country dummies	no	no	no	yes	yes	yes	yes
Period dummies	no	no	no	no	no	yes	yes
R^2 :	0.27	0.29	0.28	0.53	0.49	0.81	0.83
Adjusted R^2 :	0.25	0.26	0.24	0.44	0.37	0.71	0.72
SE:	0.017	0.017	0.018	0.015	0.016	0.011	0.011
D-W ^a :	2.31	2.47	2.57	2.19	2.25	2.14	2.01
Sample size:	44	44	38	44	38	44	38
d.f.:	42	41	35	36	32	28	22

Notes: Numbers in parentheses below the coefficient estimates are *t* statistics. Key: τ^h = ratio of country *h*'s technology level to U.S. technology level; κ^h = ratio of country *h*'s capital:labor ratio to the U.S. capital:labor ratio. Computations are based on gross capital stock (net for Germany) and national income-based factor shares averaged between the United Kingdom and the United States. For regressions 1, 2, 4, and 6, observations are for Germany, Italy, Japan, and the United Kingdom for nine periods: 1880–1890, 1890–1900, 1900–1913, 1913–1929, 1929–1938, 1938–1950, 1950–1960, 1960–1970, and 1970–1979; Canada for 1929–1938, 1938–1950, 1950–1960, 1960–1970, and 1970–1979; and France for 1950–1960, 1960–1970, and 1970–1979; for regressions 3, 5, and 7, the 1880–1890 observation is excluded.

^aDurbin-Watson statistic, based on observations ordered within country over time.

*Significant at the 5-percent level; **significant at the 1-percent level.

The sample for (7a) consists of six countries (excluding the United States) for each of nine time periods (with available data): 1880–1890, 1890–1900, 1900–1913, 1913–1929, 1929–1938, 1938–1950, 1950–1960, 1960–1970, and 1970–1979; the sample size is 44. The sample for (7b) is the same, except that the 1880–1890 observation is excluded; the sample size is 38.

Two tests are made of the interaction hypothesis. The first [specification (7a)] posits a positive association between the rate of growth of relative productivity, $\hat{\tau}^h$, and the rate of growth of the relative capital:labor ratio, $\hat{\kappa}^h$. This is consistent with the general formulation of the interaction effect. The second [specification (7b)] posits a positive relation between $\hat{\tau}_i^h$ and the change in $\hat{\kappa}^h$. The latter is consistent with the strict vintage model, since the change in

the average age of the capital stock depends on the *acceleration* in the rate of capital growth (see Nelson, 1964). Also, as is implicit in the two specifications, the catch-up hypothesis is tested (a negative coefficient on τ^h).

Results for the interaction hypothesis, shown in Table 3, are generally supportive. Coefficient estimates for relative capital:labor growth are all positive (columns 2, 4, and 6). The coefficient estimate for $\hat{\kappa}$ is not significant when no country or period dummy variables are included, significant at the 5-percent level when country dummies are included, and significant at the 1-percent level when both sets are included. The latter is perhaps the most revealing result, since it suggests that it is the residual variation in TFP growth, after country- and time-specific effects are removed, that is

most strongly correlated with the variation in capital:labor growth.

The results for $\Delta \hat{\kappa}_t^h$ (columns 3, 5, and 7) are all positive, as predicted, but insignificant except when both country and time dummy variables are added, in which case it is statistically significant at the 5-percent level. Thus, (7a) provides a better fit to the data.¹² The results also confirm the catch-up hypothesis, at the 1-percent significance level, for this (highly selective) sample of countries.

The F test for the inclusion of the country dummy variables suggests that there are country-specific effects—economic, cultural, and institutional—that play an important role in productivity growth. Dummy variables (relative to the United Kingdom) are not significant for Canada and France in equation (7a) but are statistically significant at the 1-percent level and negative for Germany, Italy, and Japan. In other words, once relative backwardness and the rate of capital:labor growth are accounted for, Germany, Italy, and Japan had lower TFP growth than the United Kingdom. My results also held for the postwar period and differ from those of Dowrick and Nguyen, who found higher than predicted TFP growth for France, Germany, and Japan during 1950–1985, once these two factors were controlled. The reason for the difference in results is not readily apparent, though Dowrick and Nguyen use a different sample of countries and base their estimate of capital stock on investment flow data.

Moreover, the F test for the inclusion of period dummy variables is significant at the 1-percent level for each specification with country dummy variables. The only two period dummies that are statistically significantly different at the 5-percent level from 1970–1979 are 1929–1938, which is positive

(a consequence of low investment rates during the Depression), and 1938–1950, which is negative (from the effects of World War II on output).

Finally, when the regressions are performed separately for prewar data (1880–1938) and postwar data (1950–1979), results are much stronger for the latter. In particular, for equation (7a), with country dummy variables, the coefficient of $\hat{\kappa}$ is 0.306, which is significant at the 1-percent level, and the R^2 statistic is 0.95 for the 1950–1979 data.¹³ For the 1880–1938 data; the coefficient of $\hat{\kappa}$ is 0.088, which is not statistically significant, and the R^2 -statistic is 0.46. An F test (or “Chow test” from Gregory Chow [1960]) on structural change between the 1880–1938 data and the 1950–1979 data is statistically significant at the 1-percent level. The results are consistent with the correlation coefficients between TFP growth and capital:labor growth by period. I will have more to say about this in the conclusion.¹⁴

¹³One might assume at first glance that the postwar results are dominated by the German and Japanese reconstruction. However, when two interactive terms, $\hat{\kappa} \times \text{DUMGER}$ and $\hat{\kappa} \times \text{DUMJAP}$, are included in the specification, $\hat{\kappa}$ remains significant at the 1-percent level.

¹⁴Additional tests were performed for both the full sample and the postwar sample. First, the observations were ordered by time within each country in order to test for autocorrelation. The Durbin-Watson statistics, shown in Table 3, all fall within the critical range (5-percent level), except one in the uncertainty range. Second, standard heteroscedasticity tests were performed for each regression equation for τ_t^h and $\hat{\kappa}_t^h$ or κ_t^h , where appropriate. The test results are all insignificant at the 5-percent level. Third, Ramsey RESET functional-form tests were performed for the square and cube of the predicted value of the dependent variable, with no significant results at the 5-percent level. For column 6 of Table 3, the $F_{[2, 39]}$ statistic is 2.65, compared to a critical value of 3.23 at the 5-percent level. Fourth, an endogeneity test was performed for $\hat{\kappa}_t^h$ or κ_t^h , where appropriate, by first regressing $\hat{\kappa}_t^h$ on initial TFP and initial capital:labor ratio or κ_t^h on initial TFP and capital:labor growth, and then including the residual from this equation in the estimating equation. For column 6 of Table 3, the t statistic for the estimated residual is 1.1 and for column 7 the t statistic is 0.9. Two other specifications were also used:

$$(7c) \quad \hat{\rho}_t^h = b_0 + b_1 \tau_t^h + b_2 \hat{\kappa}_t^h + \varepsilon_t^h$$

$$(7d) \quad \hat{\rho}_t^h = b_0 + b_1 \tau_t^h + b_2 \Delta \hat{\kappa}_t^h + \varepsilon_t^h$$

¹²The variable $\hat{\gamma}_t^h \equiv \hat{K}_t^h - \hat{K}_t^{\text{US}}$ is also used in place of $\hat{\kappa}_t^h$ in (7a) and $\Delta \hat{\gamma}_t^h$ in place of $\Delta \hat{\kappa}_t^h$ in (7b). The former is insignificant except when both country and time dummy variables are added, in which case it is statistically significant at the 5-percent level. The latter is not statistically significant in any form. The interaction effect is more strongly related to changes in capital intensity than to changes in total capital stock.

C. Vintage Capital

A third approach is to construct a vintage model of productivity growth, which relates the level of productivity not only to the level of the capital stock but also to the age distribution of the capital stock. To simplify, suppose that this year's capital investment is s percent more productive than last year's and that the parameter s is constant over time. Let \bar{A}^h be the average age of country h 's capital stock. Then,

$$(8) \quad \ln Y^h = \zeta^h + \alpha \ln L^h + (1 - \alpha) \ln K^h - (1 - \alpha) s^h \bar{A}^h.$$

The average age of the capital stock is estimated from capital-stock data for 1870, 1880, 1890, 1900, 1913, 1929, 1938, 1950, 1960, 1970, and 1979. It is assumed that the service life is 50 years and that the average age of the capital stock was 25 years in 1870. Estimates are not provided for Canada or France, because the capital-stock series are not long enough.

Results on average age are shown in Table 4. Capital-stock age moves inversely with changes in the rate of growth of the capital stock. If growth accelerates, the average age of the capital stock declines. Changes in the rate of growth of the capital stock are positively associated with the actual rates of growth of the capital stock (the correlation coefficient is 0.60). The average age for the five countries declined from 23 years in 1880 to 21 years in 1913, rose steadily to 28 years in 1950, then rapidly declined to 15 years in 1979. It is also interesting that the standard deviation of capital-stock age increased between 1880 and 1938. Over the postwar period, it remained relatively constant, except for a drop in 1960. This is true despite the convergence in capital:labor ratios after 1960.

The United States had by far the newest capital stock from 1880 to 1913 (one-third

younger than the other four countries in 1900), a consequence of its high rate of capital growth. U.S. capital stock aged relative to the other countries from 1900 onward, and by 1979 it was 13-percent older than the that of other countries. From 1929 onward, Japan had the youngest capital stock; in 1979, its average age was two-thirds that of its nearest rival, Germany, and 0.58 that of the United States. In contrast, the United Kingdom had the oldest capital stock, a position it maintained for 100 years. In fact, in 1900, the U.K. capital stock was 70-percent older than that of the United States.

From (6) and (8) and with the added assumption that s is equal across countries, it follows that

$$(9) \quad \hat{\pi}^h = \hat{\tau}^h + (1 - \alpha) \kappa^h - (1 - \alpha) s \Lambda^h$$

where $\Lambda^h \equiv d\bar{A}^h/dt - d\bar{A}^{US}/dt$, the difference in the *rate of change* in capital-stock age between country h and the United States (see Nelson, 1964). I also estimated the following regression equation:

$$(10) \quad \hat{\tau}_t^h = b_0 + b_1 \tau_t^h + b_2 \Lambda_t^h + \varepsilon_t^h.$$

The results for Λ^h are all negative, as predicted, and significant at the 1-percent level when no dummy variables are included ($R^2 = 0.41$) and when country dummy variables are included ($R^2 = 0.55$). The estimated value of s , on the basis of the first two forms, is 0.0082.¹⁵

D. Capital:Output Constancy

A constant capital:output ratio within countries can also lead to a positive interaction effect between TFP growth and the growth in the capital:labor ratio. It follows from this that the correlation between country TFP growth (ρ^h) and the growth in

These differ from (7a) and (7b) in using actual country TFP and capital:labor growth rates instead of relative rates. The sample also includes data from the United States. Results are similar to those reported in Table 3.

¹⁵The variable Λ^h is significant at only the 10-percent level when both country and time dummy variables are added, as a result of the high multicollinearity of Λ^h with time. A simple regression of Λ^h on the eight time dummy variables yields an R^2 of 0.58.

TABLE 4—AVERAGE AGE OF CAPITAL STOCK AND CAPITAL:OUTPUT RATIOS, 1870–1979

A. Average age of capital stock ^a											
Country	1870	1880	1890	1900	1913	1929	1938	1950	1960	1970	1979
Germany	—	23.0	21.9	20.4	20.0	27.3	28.7	30.9	19.4	14.4	15.3
Italy	—	24.3	22.9	24.2	21.9	22.8	21.5	25.9	21.2	16.4	15.8
Japan	—	24.6	25.9	24.3	20.5	16.9	18.4	23.4	19.6	10.7	10.0
United Kingdom	—	25.6	27.1	27.0	25.7	28.5	30.1	31.4	24.8	19.3	19.2
United States	—	19.5	18.4	15.8	16.5	20.3	25.5	26.7	22.7	19.0	17.3
Mean:	—	23.4	23.2	22.3	20.9	23.2	24.8	27.7	21.5	16.0	15.5
SD:	—	2.1	3.1	3.9	3.0	4.3	4.4	3.1	2.0	3.2	3.1
Average/U.S. ^b :	—	1.25	1.33	1.52	1.33	1.18	0.97	1.04	0.93	0.80	0.87
B. Gross capital:GDP ratio (1970 U.S. prices) ^b											
Country	1870	1880	1890	1900	1913	1929	1938	1950	1960	1970	1979
Canada	—	—	—	—	—	3.23	3.54	2.41	2.64	2.56	2.69
France	—	—	—	—	—	—	—	2.80	2.34	2.24	2.44
Germany	2.96	3.24	3.20	3.17	3.33	3.11	2.57	3.01	2.29	2.64	3.14
Italy	—	1.35	1.65	1.70	1.77	1.99	2.27	2.31	1.99	2.02	2.35
Japan	—	1.30	1.18	1.19	1.36	1.53	1.39	2.33	1.65	1.80	2.47
United Kingdom	2.56	2.49	2.25	2.18	2.28	2.29	2.15	1.97	1.92	2.16	2.42
United States	2.51	2.49	2.40	2.87	3.13	3.04	3.47	2.41	2.48	2.48	2.51
Mean: ^c	—	2.17	2.14	2.22	2.37	2.39	2.37	2.41	2.06	2.22	2.58
SD:	—	0.75	0.87	0.97	1.04	1.01	1.14	0.91	0.85	0.87	0.98
C. Average annual growth rates in capital:output ratios (percentages)											
Country	1880–1938			1950–1979				1880–1979			
Canada	—			0.38				—			
France	—			-0.47				—			
Germany	-0.40			0.15				-0.03			
Italy	0.89			0.06				0.56			
Japan	0.12			0.20				0.65			
United Kingdom	-0.25			0.71				-0.03			
United States	0.57			0.14				0.01			
Mean: ^b	0.19			0.25				0.23			
SD: ^b	0.49			0.23				0.31			

^aAverage age is estimated from capital-stock data for 1870, 1880, 1890, 1900, 1913, 1929, 1938, 1950, 1960, 1970, and 1979. It is assumed that the service life is 50 years and that the average age of the capital stock was 25 years in 1870.

^bRatio of unweighted average ages for all countries except the United States to the U.S. age.

^cThe mean is the unweighted average for Germany, Italy, Japan, the United Kingdom, and the United States; the standard deviation is based on the same five countries.

the country capital:labor ratio ($\hat{k}^h \equiv \hat{K}^h - \hat{L}^h$) will be 1.0, even if technical change is disembodied. The result follows directly from (2) that, if $\hat{Y} = \hat{K}$, then

$$(11) \quad \rho_t^h = \alpha \hat{k}_t^h$$

or, equivalently,

$$(12) \quad \hat{\tau}_t^h = \alpha \kappa_t^h.$$

There are several theoretical justifications for this position. For example, in a Solow growth model, with a Cobb-Douglas production function with disembodied technical change such as (3), then in steady-state equilibrium the capital:output ratio will be constant (see Solow, 1956).

There are three testable implications of this model. First, capital:output ratios

should be constant over time. However, results in panels B and C of Table 4 indicate that country capital:output ratios have changed over long periods, such as for Italy and the United States for 1880–1938; Canada, France, and the United Kingdom for the 1950–1979 period; and Italy and Japan over the 1880–1979 period. On average, capital:output ratios trended upward between 1880 and 1979. Moreover, the standard deviation of capital:output growth rates does indicate noticeable differences in the experiences of the five countries. However, changes in capital:output ratios have been considerably less than those for capital:labor ratios.¹⁶

Second, the correlation coefficient between TFP growth and the growth in the capital:labor ratios should be unity. The cross-country correlation coefficients were 0.08 for 1880–1913, 0.37 for 1913–1938, 0.56 for 1938–1950, 0.95 for 1950–1979, and 0.79 for 1880–1979. Thus, except for the postwar period, the correlation coefficients are much lower than this model would predict.

Third, from (12), the regression coefficient of $\hat{\kappa}$ in equation (7a) should equal the mean wage share, which for this sample was 0.598. I performed t tests for $\hat{b}_2 = 0.598$ on the basis of the three sets of regression results reported in Table 3 for the 1880–1979 period, and the nulls were rejected at the 1-percent level (one-tail test) in two cases and at the 5-percent level (one-tail test) in the third case. It was also rejected at the 1-percent level for the regression on postwar data with country dummy variables. These results suggest that the finding of a positive interaction effect cannot be simply ascribed to capital:output constancy.

E. Increasing Returns to Capital

Paul Romer (1986) has argued that increasing returns to capital and externalities

from new knowledge development may account for rising worldwide productivity growth. This argument is based on the finding that labor productivity growth has been increasing over the long run, from 1770 to 1979. Romer argued that the rising world productivity growth may be due, in part, to the increasing returns to scale. It is difficult to discriminate directly between this model and the interaction hypothesis. A simple procedure is to replace $\hat{\kappa}_t^h$ with κ_t^h , the relative level of capital intensity in country h at time t (or γ_t^h , the relative level of capital stock) in equation (7a). Country and period dummy variables are also included in alternative specifications. The estimated coefficients of κ_t^h (and γ_t^h) are all statistically insignificant. Jess Benhabib and Boyan Jovanovic (1991), using U.S. aggregate data and the Summers-Heston sample, also find little evidence to support externalities to capital in a modified Romer model.

IV. Concluding Remarks

It is illuminating to recast the results in an historical frame. In 1880, the United Kingdom was the leading nation in terms of productivity and capital intensity, and the United States was second, while Italy and Japan were still in the earliest stages of industrialization. Between 1880 and 1938, despite two major depressions and a world war, modest convergence occurred in both TFP and labor productivity among the countries in the sample. The United States had by far the highest growth in capital intensity and, by 1938, had the newest capital stock, the highest capital:labor ratio (three times the average of the other countries), and the highest TFP level (45-percent greater than the average of the other countries). The United Kingdom was last in capital:labor growth and, by 1938, had slipped to third in terms of capital intensity and second in terms of TFP.

The 1938–1950 period, dominated by World War II, saw the United States surge ahead of the rest of the countries in terms of technological leadership. By 1950, the TFP level in the United States was more than twice the average of the other coun-

¹⁶However, it should be noted that the capital:output ratios have not been adjusted for changes in capacity utilization. Such an adjustment may reduce the variability in the estimated capital:output ratio over time.

tries. Moreover, absolute declines in TFP levels were recorded for Germany and Japan.

The postwar period, from 1950 to 1979, was characterized by very strong convergence in TFP, labor productivity, and capital intensity among the group of seven. Part of this process was due to the postwar recovery of Germany and Japan. However, this was also a period characterized by historically unprecedented high rates of TFP growth, labor productivity growth, and capital:labor growth. The United States maintained its technology leadership, but its relative position dwindled from more than twice the average of the other countries to a 22-percent differential. The United States had the lowest rate of capital:labor growth, and by 1979 its capital stock was 15-percent older than the other countries' and 73-percent older than Japan's.

In summary, the emergence of the United States as the technological leader in 1900 and the widening gap between the United States and rest of the world through 1950 coincides with a very high rate of capital:labor growth, its dominance in terms of capital intensity, and its new capital vintages. Its dwindling leadership position during the postwar period is coincident with low capital:labor growth and the aging of its capital stock. The loss of technological leadership by the United Kingdom after 1890 and the almost continuous slippage in its relative position thereafter is associated with a low rate of domestic investment and the relative aging of its capital stock.

Finally, the rate of catch-up of individual-country technology levels was positively associated with the rate of growth of the capital:labor ratios. However, the strength of this association varied over time and was strongest after World War II. Indeed, the regression results with the prewar data yield a positive but statistically insignificant interaction effect. These results are consistent with the results of Abramovitz (1986), who found sluggish convergence in labor productivity before World War II and rapid convergence after. He attributed the difference to the lack of "social capability," low educational levels, and inadequate industrial and

financial organization before World War I; between 1913 and 1950, the process was interrupted by two world wars and the Great Depression; and only after 1950 was the process rapid and smooth, because all the elements for the catch-up process were in place.

A strong interaction effect appears to occur when inhibitions to growth are eliminated. Such a period is characterized by high (average) productivity growth, a high rate of capital formation, a large initial dispersion of technology levels, and rapid catch-up in technology, as characterized the years from 1950 to 1979. The interaction effect is likely to be weak when impediments to growth are present or when differences in technology levels among countries are small.

REFERENCES

- Abramovitz, Moses, "Rapid Growth Potential and its Realization: The Experience of Capitalist Economies," in Edmond Malinvaud, ed., *Economic Growth and Resources*, Proceedings of the Fifth World Congress of the International Economic Association, Vol. 1, London: Macmillan, 1979, pp. 1-30.
- _____, "Catching Up, Forging Ahead, and Falling Behind," *Journal of Economic History*, June 1986, 46, 385-406.
- _____, and David, Paul A., "Reinterpreting Economic Growth: Parables and Realities," *American Economic Review*, May 1973 (*Papers and Proceedings*), 63, 428-39.
- Arrow, Kenneth, "The Economic Implications of Learning by Doing," *Review of Economic Studies*, June 1962, 29, 155-73.
- Baumol, William J., "Productivity Growth, Convergence, and Welfare: What the Long-Run Data Show?" *American Economic Review*, December 1986, 76, 1072-85.
- _____, and Wolff, Edward N., "Productivity Growth, Convergence, and Welfare: Reply," *American Economic Review*, December 1988, 78, 1155-9.
- Benhabib, Jess and Jovanovic, Boyan, "Externalities and Growth Accounting," *Ameri-*

- can Economic Review*, March 1991, 81, 82-113.
- Bombach, Gottfried, *Post-War Economic Growth Revisited*, Amsterdam: North-Holland, 1985.
- Chow, Gregory C., "Tests of Equality between Sets of Coefficients in Two Linear Regressions," *Econometrica*, July 1960, 28, 591-605.
- Coe, David and Holtham, Gerald, "Output Responsiveness and Inflation: An Aggregate Study," *OECD Economic Studies*, Autumn 1983, 1, 94-145.
- Deane, Phyllis and Cole, W. A., *British Economic Growth, 1688-1959: Trends and Structure*, Cambridge: Cambridge University Press, 1964.
- De Long, Bradford, "Productivity Growth, Convergence, and Welfare: Comment," *American Economic Review*, December 1988, 78, 1138-54.
- Dowrick, Steve and Nguyen, Duc-Tho, "OECD Comparative Economic Growth 1950-85: Catch-Up and Convergence," *American Economic Review*, December 1989, 79, 1010-31.
- Gerschenkron, Alexander, "Economic Backwardness in Historical Perspective," in Bert F. Hoselitz, ed., *The Progress of Underdeveloped Areas*, Chicago: University of Chicago Press, 1952, pp. 3-29.
- Goldsmith, Raymond W., *Comparative National Balance Sheets* Chicago: University of Chicago Press, 1985.
- Gollop, Frank M. and Jorgensen, Dale W., "U.S. Productivity Growth by Industry, 1947-73," in John W. Kendrick and Beatrice N. Vaccara, eds., *New Developments in Productivity Measurement and Analysis*, Chicago: University of Chicago Press, 1980, pp. 17-124.
- Johnson, D. Gale, "The Functional Distribution of Income in the United States, 1850-1952," *Review of Economics and Statistics*, May 1954, 36, 175-82.
- Kendrick, J. and Sato, R., "Factor Prices, Productivity, and Economic Growth," *American Economic Review*, December 1963, 53, 974-1003.
- Kuznets, Simon, *Population, Capital, and Growth: Selected Essays*, New York: Norton, 1973.
- Leamer, Edward E., *Sources of International Comparative Advantage*, Cambridge, MA: MIT Press, 1984.
- Maddison, Angus, *Phases of Capitalist Development*, Oxford: Oxford University Press, 1982.
- , "Growth and Slowdown in Advanced Capitalist Economies: Techniques of Quantitative Assessment," *Journal of Economic Literature*, June 1987, 25, 649-706.
- Nelson, Richard R., "Aggregate Production Functions and Medium-Range Growth Projections," *American Economic Review*, September 1964, 54, 575-605.
- Ohkawa, Kasushi and Rosovsky, Henry, *Japanese Economic Growth: Trend Acceleration in the Twentieth Century*, Stanford, CA: Stanford University Press, 1973.
- Romer, Paul M., "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, October 1986, 94, 1002-37.
- Solow, Robert M., "Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics*, February 1956, 70, 65-94.
- , "Growth Theory and After," *American Economic Review*, June 1988, 78, 307-17.
- Summers, Robert and Heston, Alan, "A New Set of International Comparisons of Real Product and Prices: Estimates for 130 Countries, 1950-1985," *Review of Income and Wealth*, March 1988, 34, 1-26.
- International Labor Organization, *Yearbook of Labor Statistics*, Geneva: International Labour Office, various issues.
- OECD Statistical Division, *Main Economic Indicators, 1960-1979*, Paris: Organization for Economic Cooperation and Development, June 1980.
- United Nations, *Yearbook of National Accounts Statistics*, New York: United Nations, various years.

Economic Forecast Evaluation: Profits Versus The Conventional Error Measures

By GORDON LEITCH AND J. ERNEST TANNER*

Economists are often puzzled as to why profit-maximizing firms buy professional forecasts when statistics such as the root-mean-squared error or the mean absolute error often indicate that a naive model will forecast about as well. This paper argues that the reason is that these traditional summary statistics may not be closely related to a forecast's profits. Using profit measures, we find only very weak relationships between such summary error statistics and forecast value. If these results are robust, then least-squares regression analysis may not be appropriate for many studies of economic behavior. (JEL C52, C10)

Economists are often puzzled as to why seemingly profit-maximizing firms buy economic forecasts. Summary statistics such as the average absolute error (AAE), the root-mean-squared error (RMSE), and the Theil "U" statistic (which is free from the dimension problems of the other two) rarely reveal major differences between professional forecasting services and a simple naive approach of no change in the variable being forecast.¹ Yet, millions of dollars are spent annually both producing and purchasing these apparently worthless forecasts.

In this paper, we argue that the conventional criteria used in these evaluations may

well be inappropriate. In empirical tests of interest rate forecasts for the 1980's, we find that the conventional criteria, based upon some measure of the size of the forecast error, have no systematic relationship to profits. Perhaps because of some embarrassing forecasts and generally poor performance on the basis of conventional criteria, few forecasters are willing to allow extensive evaluations of their historical forecasts of interest rates.² However, we did have access to the entire record of one professional service and have used it as a basis for comparison. Moreover, with interest rate forecasts, a profit measure can easily be calculated, thereby permitting tests of the relationship between profits and the conventional measures of forecast-error magnitudes. It is naturally better to examine profits directly than to examine a proxy that is at

*Graduate student and Professor, respectively, Department of Economics, Tulane University, New Orleans, LA 70118. We are grateful to Vittorio Bonomo for many insightful discussions in evaluating economic forecasts. In addition, Ron Batchelder, John Boschen, Tom Mayer, Michael Parkin, Ed Tower, Terry Wilford, Jeff Zabel and three anonymous referees offered many helpful comments on earlier drafts. However, they are not responsible for any of the paper's shortcomings.

¹Among those who accept the null hypothesis of no value added by most sophisticated forecasts over simple ARIMA-type models are Charles R. Nelson (1972), J. P. Cooper and Nelson (1975), G. V. L. Narashimhan (1975), J. R. Schmidt (1979), A. C. Petto (1981), and David Ahlers and Josef Lakonishok (1983). However, some evaluators do find that the larger models do supply valuable information not contained in the simple forecasts. See, for example, the evidence contained in Carl F. Christ (1975), Stephen K. McNees (1975, 1979), Roger Craine and Arthur M. Havenner (1988), and E. Philip Howrey et al. (1974).

²The semiannual *Wall Street Journal* survey occasionally makes mention of these problems. For example, in the July 1987 survey, it was noted that the strong economy produced sharply higher bond yields during early 1987, and many forecasts "proved widely inaccurate during the first half." A possible consequence of this was that one forecaster bowed out after predicting "anemic economic growth ... as well as a sharp decline in bond yields (July 6, 1987, p. 2) The *Blue Chip Economic Indicators* service has experienced similar difficulties as many of their interest rate forecasters do not permit identification, while all of their GNP forecasters are explicitly identified next to their forecasts.

best indirectly related to profits. Economists generally assume that firms use forecasts because they add to profits. Thus, a more appropriate test of forecast accuracy is profitability, and not the size of the forecast error or its squared value.

The paper proceeds as follows. Section I reviews the commonly used criteria to evaluate forecasts, while Section II describes a number of ways to calculate profit measures of an interest rate forecast. In Section III, we describe the interest rate forecasting approaches used in our evaluations, including the professional service and the more naive approaches. In Sections IV and V, we evaluate the forecasts in terms of the alternative profitability measures described in Section II and in terms of the conventional error-measures criteria widely used in the literature. In Section VI, we summarize the paper and offer some tentative conclusions.

I. Widely Used Forecast-Evaluation Criteria

Economic forecasts are typically evaluated by comparing the errors obtained when measuring the forecast values against the actual outcomes. The three commonly used statistics evaluating forecasts in this manner are the average absolute error (AAE), the root-mean-squared error (RMSE), and the Theil "U" coefficient.

The more familiar AAE and RMSE criteria can be misleading in certain cases. For example, a forecaster using the unit of measurement of thousands of dollars will have different error values than another forecaster using millions of dollars. Theil's U statistic is free of these problems. It can be viewed as the RMSE of a forecast divided by the RMSE of a naive forecast of no change:

$$(1) \quad U = \sqrt{\frac{\sum_{t=1}^N (\Delta F_{t+j} - \Delta A_{t+j})^2}{\sum_{t=1}^N (\Delta A_{t+j})^2}}$$

where F_{t+j} is the forecast value j periods ahead, A_{t+j} is the actual realized value of a

variable, N is the number of forecasts, $\Delta F_{t+j} = F_{t+j} - A_t$, and $\Delta A_{t+j} = A_{t+j} - A_t$. $U = 0.0$ when the prediction is perfect, $U = 1.0$ when the RMSE of the predicted change equals the accuracy of the baseline forecast of no change, and statistics greater than 1.0 indicate that the forecasts have higher RMSE's than the no-change forecast. Because of the appeal of least squares based on a quadratic loss function, in which larger errors receive proportionately more weight, the RMSE and the Theil U coefficient have been the most popular criteria in the literature.³ Other criteria have been proposed, but cataloging them would be a distraction.

II. Profit Measures of Using an Interest Rate Forecast

While there are many ways to calculate profits for an interest rate forecast, most alternatives cannot be studied, because the data are not available. For example, while the professional service's forecasts that we studied were issued monthly for 14 interest rates covering each of the succeeding 12 months,⁴ few of these interest rates are represented by a corresponding forward rate and a futures contract. An exception, the three-month Treasury bill rate, has received the most attention in the literature. Since there is a liquid futures market for this security, we analyzed only this rate in our

³We know of one prominent forecaster who purposely was pessimistically biased in his forecasts. Because he believed that the profit-maximizing firms wanted to protect themselves from the "worst-case scenario," his forecasts were purposely biased. Clearly, since he believed that his clients' loss functions were not symmetrical, the conventional criteria would not be appropriate for judging the accuracy of his forecasts. The implications of this forecaster's beliefs have much more significance, for if he is correct, then least-squares regression analysis may not be appropriate for many empirical studies of economic behavior.

⁴The monthly forecasts on 14 interest rates for the period 1-12 months ahead are done by the Commonwealth Research Group and released through the *Money Rate Report* on the final trading day of each month. We should note that one of the authors is a cofounder of this firm and retains a financial interest in it.

tests. Moreover, this is one of the interest rates for which we had access to a monthly survey of money managers, economists, corporate treasurers, and other forecasters up to a year ahead of each of the futures contract months (March, June, September, and December).

Even narrowing the available data set to this extent leaves many measures of profit available. Are profits calculated from the cash market, the futures market, or the forward market? What is the size of the position assumed, and does it change depending upon the forecast? What are the transaction costs including brokers' fees and bid-ask spread?

To simplify matters, but to preserve the maximum level of realism corresponding to that for a user of interest rate forecasts, we did our profit calculations in the futures market for Treasury bills. We assume that the size of the position is always one unit;⁵ all transactions were executed at the closing price at the end of the month,⁶ when the

professional service's forecasts became available; and the positions were always evaluated at one-month intervals. New positions, based upon the updated forecasts, were assumed to have been taken at the closing price of each month.

For trading in Treasury-bill futures, round-trip broker's fees would range from about \$8.50 each for large traders to about \$85.00 for full-service retail brokers. The initial margin requirement for trading the \$1 million Treasury-bill contract is set by the broker and usually ranges between \$2,000 and \$2,500, but falls to \$1,500 at most brokers for maintenance purposes. To deal in such a contract, the usual procedure is to deposit with the broker between \$10,000 and \$20,000 in a money-market fund earning a competitive short-term interest rate. As one trades, funds are removed from and deposited to the trading account as needed. Thus, for most practical purposes, the \$2,000–\$2,500 margin earns no interest while the rest of the funds earn a competitive rate of return. If the market moves unfavorably, funds are transferred from the money-market fund to the trading account in order to meet the maintenance margin. However, should the market move in a favorable direction, excess funds will be pulled from the trading account and placed in the money-market fund earning additional interest. No interest will be earned on the maintenance margin required to hold the position in a "speculation" account. However, the same is not true for hedgers who have lower margin requirements.

In short, if a nonhedger uses a discount broker and the market fluctuates randomly, then round-trip costs amount to about \$40 per month (\$25 broker's fees and \$15 lost interest on \$2,000) to speculate on the price change of a single \$1 million Treasury-bill futures contract. In our tests, and in actual practice, the transactions costs would be less, because we assumed new trades only when the new forecast calls for one to reverse positions or to roll over from an expiring contract.

In this environment, there are still many ways to use an interest rate forecast. Our approach for the first profit calculation

⁵Vittorio Bonomo (1989) finds that professional commodity-market traders do not vary the size of their bet, but amateurs, who end up as losers, do vary the size of their positions. He argues that if the odds are in one's favor, a series of constant-sized bets virtually guarantees success, and the professionals profit by making a large number of bets so that the odds work for them. On the other hand, traders who lose their money tend to assume different-sized bets over time. This result is obvious if one thinks of "doubling up" when the odds get high. In this case, unless the forecast is certain, such a strategy eventually always would result in losing money. For counterexamples to Bonomo's assertions, see Richard A. Epstein (1977 Ch. 3). In the context of a log utility function, a strategy of proportional betting, in which the size of the bet is proportional to expected profits, could dominate the constant-bet strategy we use. Such a strategy may make profits more highly correlated with the RMSE criterion.

⁶For most traders, placing a buy or sell "on-close" order would not have any appreciable effect on price. Given our assumption of one contract, it is very unlikely that our profit calculations would be materially affected by the effect such an order would have on this relatively fluid market. In fact, our broker at Shearson says there is a 99-percent probability of a fill at this price as several thousand contracts were traded on a typical day during the sample period used. Thus, the assumption that fills are done at closing prices gives the most accurate *ex post* prices available.

(profit-rule A) was straightforward: if interest rates are forecast to rise, go "short" or sell a futures contract; conversely, if rates are forecast to fall, go "long" or buy a futures contract. This approach implicitly assumes that the market does not expect interest rates to change. On a \$1 million three-month Treasury bill, each basis-point change in the interest rate changes the gross return by \$25 ($\$1 \text{ million} \times 0.01 \text{ percent} \times 0.25 \text{ years}$). Thus, for each basis-point change in interest rates on the futures contract, the price of the contract changes by \$25. Gross returns, before costs, are calculated directly using \$25 per basis-point change and the position taken. A \$15 opportunity cost on the margin funds per month and \$25 per round-trip trade were then subtracted to obtain trading profits.

Our second profit calculation (profit-rule B) is similar but takes the futures-market forecast explicitly into account. This profit calculation assumes that, if interest rates are forecast to be above the rate implied by the futures market, one will short the contract; and conversely, if rates are expected to be below the rate implied by the futures contract, one will go long the contract. After this position is taken, profits are calculated in the same manner as above.

Our third profit calculation (profit-rule C) assumes a position only if interest rates are expected to change. If the forecast is for no change in rates (saying one doesn't know what to do), then the action is not to take a position. Otherwise, positions are assumed as in the second profit calculation above by comparing the forecast with the market's forecast.

Our fourth profit calculation (profit-rule D) assumes a position only if the forecast change in interest rates is opposite in sign to that of the market's forecast change. In other words, we take a long position in the futures contract only if we forecast rates to fall while the futures market is forecasting a rise, and conversely. At other times, no position is taken.

Because the professional interest-rate-forecasting service offers a prediction of the three-month Treasury bill rate for each of the next 12 months, we could theoretically

estimate profits for each of the 12 forecast horizons. However, because we wanted to use the Treasury-bill forward market as an alternative forecasting system, we used only the forecasts up to nine months ahead. This is because the longest-maturing cash Treasury bill is only for one year. As a result, the longest readily estimated forward rate for three-month Treasury bills is nine months.

In predicting ahead nine months for the forecasts, we matched the forecast month with the nearest contract that would not expire before the end of the month (e.g., the December, January, and February forecasts were matched with the March futures contract; the March, April and May forecasts were matched with the June contract). Depending upon the current month, this means that the nearest three or the nearest four contracts are traded.

III. Forecasting Systems

For the tests made in this paper, we employed seven different forecasting systems. The first forecasting system was based upon the professional service whose forecasts were published in the *Money Rate Report*. Although the complete set of forecasts started in 1980, we only used the period beginning with their year-end 1981 forecast. At that point, the service began consistently including executive survey forecasts for interest rates. The survey, usually done during the last 10 days of the month, provides a "consensus" forecast for the three-month Treasury-bill rate for each month of the futures contracts up to a year ahead and fits well with our tests. Upon introducing the survey, the service indicated that the survey respondents included "[c]orporate financial officers, banking executives, and operating heads of firms ... [because] many analysts feel that market derived forecasts do not fully nor accurately represent informed opinion as to the future course of interest rates ..." (*Money Rate Report*, January 1981, p. 1). Because the survey results for the three-month Treasury bill rate were available monthly only since year-end 1981, we use the sample period from January 1982 through December 1987 for our tests.

The other forecasting systems are more conventional. An ARIMA model was estimated initially over the seven-year period beginning in January 1975 and ending in December 1981. The structure of the one-month change in three-month Treasury-bill interest rate appeared to be of the form AR(2). The fitted initial equation was estimated to be

$$(2) \quad r_t = 0.337 + 1.220 \text{ 3MoTBill}_{t-1} \\ (1.00) \quad (10.16) \\ -0.255 \text{ 3MoTBill}_{t-2} \\ (2.06)$$

($R^2 = 0.886$). For the purposes of the tests in this paper, we used a rolling regression with a fixed beginning point but ending with the month in which the above forecasts were made. Thus, aside from the general structure, only information available in the marketplace at the time the positions were assumed to have been taken was used in our *ex post* ARIMA forecasts. The final one-month-ahead ARIMA-forecast equation, estimated from January 1975 until December 1987 and looking very similar to the initial equation, was

$$(3) \quad r_t = 0.387 + 1.133 \text{ 3MoTBill}_{t-1} \\ (1.91) \quad (14.25) \\ -0.181 \text{ 3MoTBill}_{t-2} \\ (2.27)$$

($R^2 = 0.918$). This methodology was followed for the other forecast horizons using the ARIMA technique.

For the forward rate, we used the end-of-month Treasury-bill yield curve, as reported in the *Wall Street Journal*. The implicit rate up to nine months ahead for the three-month Treasury-bill rate was estimated from the Treasury-bill yield curve's bid price using the appropriate maturity dates up to one year ahead.

The futures rate forecasts were derived from the historical prices of the four nearest Treasury-bill future contracts. For forecasts nearer than the nearest contract, we interpolated between the current spot rate and the rate implied by the nearest contract.

For forecast dates between contract dates, we interpolated between the implied forecasts of the adjoining contract dates.

The other two forecasting systems are the naivest of models. The "naive no-change" technique forecasts that all future rates will equal the current spot rate. The "constant rate of growth" technique forecasts that the change in interest rates over the next x months is compounded from the most recent one-month change. Thus, rising interest rates are expected to continue to rise, while falling rates are expected to continue to fall.

IV. A First Look at the Data for Economic Forecasts: Profits and Other Criteria

In most evaluations of economic forecasts, profits are totally ignored. In Table 1, we present summary figures on the six-year average for the nine forecast horizons using the six forecasting techniques. The profitability of the forecast appears to bear little relationship to the conventional size-of-error criteria.

For the purpose of the table, we assumed that any forecast for interest rates to rise was associated with selling the appropriate futures contract and any forecast for interest rates to fall involved buying the contract. The no-change forecast involved sitting on the sidelines. Thus, the naive no-change forecast produces no profits or losses. As described in Section II, profits were calculated using the four nearby contracts corresponding to interest rate forecasts up to nine months ahead. We will show later that the lack of a clear relationship between profits and the conventional criteria is not due to the way we calculated profits, as all profit-calculation methods produced similar results. First, we need to explain what we did more fully.

The column labeled "average directional accuracy" shows the percentage of interest rate changes in the futures market that were accurately forecast by each technique over the one-month observation interval until the new forecasts were available. The results are not surprising because the forecasts are predicting revisions in what is essentially the market's full-information forecast. If the

TABLE 1—PROFITS AND OTHER FORECAST-EVALUATION CRITERIA, JANUARY 1982–DECEMBER 1987

Forecasting technique	Six-year average annual profits	Average directional accuracy (percentage)	Average absolute error (percentage)	Average root-mean-squared error (percentage)	Average Theil <i>U</i> statistic
Professional service	\$1,643	49.3	0.781	0.932	1.93
ARIMA	–\$928	47.9	0.739	0.902	1.82
Forward rate	–\$3,050	43.7	0.656	0.848	1.62
Naive no-change	\$0	37.9	0.410	0.530	1.00
Constant rate of growth	–\$674	46.6	2.013	2.514	4.72
Survey forecast	–\$3,262	43.8	0.811	1.081	2.06

Profit Rule

Buy a \$1 million Treasury-bill futures contract if interest rates are forecast to fall, and sell a contract if rates are expected to rise. Profits are calculated based on a \$25-per-basis-point change in the futures contract over a one-month period, until a new forecast is made at the close of each month.

Note: Although the level of profits, directional accuracy, and size of forecast error are different depending on the profit rule, the general pattern is very similar across trading rules, and in no case does there seem to be a strong and theoretically correct relation between profits and the error criteria. Thus, for brevity, we summarize the data using profit-rule A, as it seems to be the most straightforward.

interest-rate market is efficient, which most studies find it to be, then predicting changes in interest rates by any system should be equivalent to flipping a coin. The possible exception is the naive no-change forecast, which predicts that the futures market rate will move toward the spot rate. The fact that it appears to forecast the wrong direction almost two-thirds of the time suggests that the spot rate probably moves to the futures market, rather than conversely, but that is not the subject of this paper.⁷

The “average absolute error” column is the average difference between the forecast value as made at the end of the previous month and the futures-market value at the end of the current month for the forecast month in question. Note that, instead of the actual realized interest rate, we have used only the next month’s new futures-market forecasts for the realization in this table. However, the results are not affected by doing the calculations this way, as the cash markets give essentially identical results. The root-mean-squared error and the Theil *U* statistics are calculated from the same

month-end data of the futures market and forecasts.

The data presented in this table are consistent with previous research. The only reference to a profit criterion that we find in the literature suggests that profits do not appear to be related to lowest root-mean-squared error. In Stephen Figlewski and Thomas Ulrich (1983), profits of a Treasury-bill forecast, when corrected for bias, are directly related to the size of the forecast error; but when they are not corrected for bias, profits and error measures are inversely related. However, the authors do not draw attention to this aspect of their research. Similarly, Scott Hein and Raymond Spudeck (1988) find no relation between cost and the conventional error criteria; but since they are concerned with using forecasts to minimize interest borrowing costs, they ignore the implications for forecast evaluation.

The columns on the size of the forecast error are also consistent with previous research; that is, in general, the smaller the forecast error, the better the Theil statistic. Moreover, as in the findings of James Pesando (1978), the naive forecast of no change has the smallest error, followed by predictions based on the forward rate and the ARIMA model. The worst forecasts, ranked from worst to best in terms of error magnitudes, are the constant-interest-rate-change forecasts followed by the survey forecasts, and then the professional-service forecasts.

⁷Michael T. Belongia (1987) found similar results for the cash market in Treasury bills. Using the results of the professional forecasters surveyed semiannually by the *Wall Street Journal*, he found they had an accuracy rate of 42 percent in predicting the direction of change over a six-month horizon. The comparable rate for the futures market was 55 percent.

TABLE 2—CORRELATIONS BETWEEN PROFITS AND THE VARIOUS FORECAST CRITERIA FOR TREASURY-BILL FUTURES MARKET TRADING AVERAGED OVER FORECAST HORIZONS OF 1–9 MONTHS

Profit rule	DA		AAE		RMSE		Theil <i>U</i>	
	<i>r</i>	<i>t</i>	<i>r</i>	<i>t</i>	<i>r</i>	<i>t</i>	<i>r</i>	<i>t</i>
A	+0.428	2.76	–0.077	0.45	–0.094	0.55	–0.156	0.92
B	+0.842	9.11	+0.185	1.10	+0.210	1.25	–0.230	1.38
C	+0.598	4.35	+0.226	1.35	+0.253	1.52	–0.148	0.87
D	+0.578	4.13	+0.326	2.01	+0.322	1.99	+0.009	0.05

Profit Rules

- A: forecast rates to rise, short the futures contract; forecast rates to fall, buy the futures contract
 B: forecast rates to be above implied forecast of futures, go short; forecast rates to be below implied forecast of futures, go long.
 C: same as B, except no position if forecast is for no change in rates
 D: take position only if forecast change has opposite directional sign to market forecast.

Notes: DA = directional accuracy of forecast; AAE = average absolute error of forecast; RMSE = root-mean-squared error of forecast; Theil *U* = Theil's forecast-evaluation statistic; *r* = simple correlation coefficient; and *t* = absolute value of the *t* statistic on the slope coefficient. For all analyses, *N* = 36 (forecasts for all nine forecast horizons were grouped together in each year, yielding six years of data on six forecasting systems, or 36 observations per profit rule).

Based upon the usual error-magnitude criteria, the results of this table confirm the major conclusions of the literature that firms appear to waste money on economic forecasts.

However, if profits were the criterion for whether or not to purchase forecasts, then the conclusion might be different. In the case summarized in Table 1, the only profitable system finished fourth out of six, based on the conventional criteria. There is nothing in these "error" criteria that would have caused firms to choose this profitable forecasting technique over the others. Yet, since the profitable forecasts were available only at a price while the others were essentially free, the market test did indicate that this forecast service's clients were probably not using the economists' criteria of lowest mean absolute error to evaluate the forecasts.

V. Statistical Relationships Between Profits and Error Measures

Although these results, which were based on six-year averages for one relatively arbitrary use of interest rate forecasts, are interesting and informative, are they statistically

significant, and do they hold up for other plausible ways of using interest rate forecasts? To answer these questions, we calculated correlations between the various criteria and profits as measured by each of the profit rules discussed in Section II.

In Table 2, we present the results of averaging each forecast system over all nine forecast horizons for each year. Thus, the data are based on 36 observations: six forecasting systems for six years. In Table 3, we present the results for each of the nine forecast horizons. Since the data are still averaged over each year, we have 324 observations, corresponding to the six forecast systems and nine forecast horizons over six years.

As Table 2 shows, regardless of the profit rule followed, there is little systematic relationship between profits and the conventional measures of forecast quality. The only conventional measure of forecast quality that is related to profits is "directional accuracy" (DA), and it is infrequently used.⁸

⁸To the best of our knowledge, only James Cicarelli (1982) and E. Philip Howrey et al. (1974) use direc-

TABLE 3—CORRELATIONS BETWEEN PROFITS AND VARIOUS FORECAST CRITERIA FOR TREASURY-BILL FUTURES-MARKET TRADING AVERAGED OVER EACH YEAR BUT WITH EACH OF THE NINE FORECAST HORIZONS TREATED SEPARATELY

Profit rule	DA		AAE		RMSE		Theil <i>U</i>	
	<i>r</i>	<i>t</i>	<i>r</i>	<i>t</i>	<i>r</i>	<i>t</i>	<i>r</i>	<i>t</i>
A	+0.441	8.81	-0.095	1.72	-0.101	1.82	-0.144	2.60
B	+0.819	25.66	+0.074	1.34	+0.096	1.72	-0.202	3.70
C	+0.619	14.14	+0.100	1.80	+0.123	2.23	-0.148	2.69
D	+0.572	12.51	+0.212	3.90	+0.216	3.97	-0.038	0.68

Profit Rules

- A: forecast rates to rise, short the futures contract; forecast rates to fall, buy the futures contract.
 B: forecast rates to be above implied forecast of futures, go short; forecast rates to be below implied forecast of futures, go long.
 C: same as B, except no position if forecast is for no change in rates.
 D: take position only if forecast change has opposite directional sign to market forecast.

Notes: DA = directional accuracy of forecast; AAE = average absolute error of forecast; RMSE = root-mean-squared error of forecast; Theil *U* = Theil's forecast-evaluation statistic; *r* = simple correlation coefficient; and *t* = absolute value of the *t* statistic on the slope coefficient. For all analyses, *N* = 324 (six forecasting systems, nine horizons, and six years of data).

The standard criteria show no consistent relationship with profits. Indeed, the average-absolute-error and root-mean-squared-error criteria have perverse signs (better forecasts should have lower average errors and higher profits, and thus, the simple correlations should be negative) in three-quarters of the cases. While the Theil statistic has the correct sign in three cases out of four, it is never statistically significant at conventional levels of significance.

Increasing the data set to 324 observations by treating each forecast horizon separately, but still averaging the data over the year, produces similar results (Table 3). However, the sizes of the error criteria are now occasionally significantly related to profits. However, of the six statistically significant correlations, three have incorrect signs. Nevertheless, directional accuracy re-

mains highly significant as a proxy for forecast profits.

In Table 4, we present the correlations and related absolute *t* values between the conventional criteria using profit rule B. For brevity, the other profit rules have been omitted, but the results are very similar. As we would expect based upon the earlier tables, directional accuracy and the error-measure criteria are only weakly related. However, the various error-measure criteria are all closely related, confirming the usual result that the accuracy of the forecast is not likely to be dependent upon which error criterion is used. Unfortunately, none of the error criteria is reliably related to the profitability of the forecast.⁹

tional accuracy in evaluating forecasts. While Cicarelli was proposing the criterion, rather than extensively evaluating forecasts, Howrey et al. were using it to bolster the record of the Wharton forecasts over the 1969–1970 recession.

⁹Because many previous studies used professional forecasts in comparisons with the ARIMA-type forecast, we have followed their example. However, in the present case, it is conceivable that our results depend upon some fortunate outliers of the particular professional forecasting service we used. To test this conjecture, we redid the tests excluding the forecasts of the professional service. In this reduced sample, all the correlation coefficients and all the *t* values were extremely close to those reported in Tables 2, 3, and 4.

TABLE 4—CORRELATIONS AMONG VARIOUS FORECAST CRITERIA

N	Statistic	Correlation					
		DA-AAE	DA-RMSE	DA-Theil U	AAE-RMSE	AAE-Theil U	RMSE-Theil U
36	<i>r</i>	+0.061	+0.077	-0.129	+0.997	+0.675	+0.657
	<i>t</i>	0.36	0.45	0.76	71.83	5.33	5.08
324	<i>r</i>	+0.012	+0.024	-0.104	+0.996	+0.744	+0.726
	<i>t</i>	0.22	0.42	1.88	202.7	19.96	18.92

Notes: DA = directional accuracy of forecast; AAE = average absolute error of forecast; RMSE = root-mean-squared error of forecast; Theil U = Theil's forecast-evaluation statistic; *r* = simple correlation coefficient; and *t* = absolute value of the *t* statistic on the slope coefficient.

Since the forecasts we have used pertain to the cash market, conclusions based on the futures market about the reliability of conventional forecast-evaluation criteria may be flawed, because the two markets do not track exactly. As a consequence, we redid the tests using the cash market. In doing so, we were forced to make some very severe and unrealistic assumptions. First was the obvious problem that any three-month Treasury bill automatically turns into a two-month Treasury bill over the one-month observation interval. To overcome this problem, we assumed a perpetual three-month Treasury bill. With this theoretical construct, profits were calculated by multiplying \$25 by the change in the three-month bill rate. Second, opportunity and trading costs in the cash market can differ widely, depending upon the characteristics of the trader, such as the bid-ask spread or the "cost of carry." To avoid these problems, we assumed trading costs identical to those applicable to futures trading with all transactions being made at the "bid" price.

While these assumptions are limiting, they do allow an operational method with which to calculate profits in the cash market, and the results are essentially identical to those derived from using the futures market. Profits and the directional accuracy of the forecast systems are highly correlated with a high degree of statistical significance. However, the relationships between profits (or directional change) and the conventional size-of-forecast-error criteria appear to be

very poorly related. In just over half the experiments tried, the relationships between profits and the root-mean-squared error have the anticipated correct negative sign; but individually, the level of significance is higher for the theoretically incorrect positive signs than for the correct negative signs. Like the futures-market evaluations, only the Theil *U* statistics generally have the theoretically appropriate signed relationship to profits, but often the level of significance is less than satisfactory as a reliable forecast-evaluation criterion.

VI. A Tentative Conclusion

Forecast evaluations made on the basis of conventional error-magnitude criteria often find little justification for profit-maximizing firms to allocate resources to professional forecasters, as naive models often predict as well. Unfortunately, the size of the conventional forecast-error criteria frequently has unpredictable relationships to the forecasts' profitability, rendering these criteria unreliable indicators of profits. Indeed, in evaluating the quality of interest rate forecasts, where forecast profits are readily calculated, we find no systematic relationship between the widely used *ex post* error criteria and *ex post* profits. The only substitute criterion for profits found in the literature that appears to be closely related is directional accuracy. The relationship between directional accuracy and profits appears to be almost as close as the relationships of the

various error criteria are to each other. However, because the root-mean-squared error, the average absolute error, and the Theil U statistic appear to be tenuously linked to profits and directional accuracy, it is not surprising that profit-maximizing firms buy forecasts in spite of their seemingly large forecast errors.

The results of this paper suggest that, if profits are not observable, directional accuracy of the forecasts might be used as the evaluation criterion. All the conventional forecast-error-magnitude criteria are only marginally related to profitability, while directional accuracy consistently demonstrates a high degree of statistical association.

A more disturbing implication of our findings is that conventional least squares may not be the appropriate estimation technique for economic behavior. If profits are not related to the size of the error, then it may be that our empirical estimates of economic relationships should not be based upon a squared error loss function.

REFERENCES

- Ahlers, David and Lakonishok, Josef, "A Study of Economists' Consensus Forecasts," *Management Science*, October 1983, 29, 1113-25.
- Belongia, Michael T., "Predicting Interest Rates: A Comparison of Professional and Market-Based Forecasts," *Review* (Federal Reserve Bank of St. Louis), March 1987, 69, 9-15.
- Bonomo, Vittorio, "The Behavior of Commodity Traders: Winners and Losers," mimeo, Virginia Polytechnic Institute and State University, September 1989.
- Christ, Carl F., "Judging the Performance of Econometric Models of the U.S. Economy," *International Economic Review*, February 1975, 16, 54-74.
- Ciccarelli, James, "A New Method of Evaluating The Accuracy of Economic Forecasts," *Journal of Macroeconomics*, Fall 1982, 4, 469-75.
- Cooper, J. Phillip and Nelson, Charles R., "The Ex Ante Prediction Performance of the St. Louis and FRB-MIT-PENN Econometric Models and Some Results on Composite Predictors," *Journal of Money Credit and Banking*, February 1975, 7, 1-32.
- Craine, Roger and Havenner, Arthur M., "Forecast Comparisons of Four Models of U.S. Interest Rates," *Journal of Forecasting*, January-March 1988, 7, 21-9.
- Epstein, Richard A., *The Theory of Gambling and Statistical Logic*, New York: Academic Press, 1977.
- Figlewski, Stephen and Urich, Thomas, "Optimal Aggregation of Money Supply Forecasts: Accuracy, Profitability, and Market Efficiency," *Journal of Finance*, June 1983, 38, 695-710.
- Hein, Scott E. and Spudeck, Raymond E., "Forecasting the Daily Federal Funds Rate," *International Journal of Forecasting*, 1988, 4 (4), 581-91.
- Howrey, E. Philip, Klein, Lawrence R. and McCarthy, Michael D., "Notes on Testing the Predictive Performance of Econometric Models," *International Economic Review*, June 1974, 15, 366-83.
- McNees, Stephen K., "An Evaluation of Economic Forecasts," *New England Economic Review* (Federal Reserve Bank of Boston), November/December 1975, 3-39.
- , "The Forecasting Record for the 1970's," *New England Economic Review* (Federal Reserve Bank of Boston), September/October 1979, 33-53.
- Narashimhan, G. V. L., "A Comparison of Predictive Performance of Alternative Forecasting Techniques: Time Series versus Econometric Models," *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, August 1975, 459-64.
- Nelson, Charles R., "The Prediction Performance of the FRB-MIT-PENN Model of the U.S. Economy," *American Economic Review*, December 1972, 62, 902-17.
- Pesando, James A., "On the Efficiency of the Bond Market: Some Canadian Evidence," *Journal of Political Economy*, December 1978, 86, 1057-76.
- Petto, A. C., "Comparing the Relative Accuracy of Seasonal Sales Forecasting Models—A Case Study of an Urban Hotel,"

unpublished paper presented at the First International Symposium on Forecasting, Quebec, Canada, May 27-29, 1981.

Schmidt, J. R., "Forecasting State Retail Sales: Econometric Versus Time Series Models," *Annals of Regional Science*,

November 1979, 13, 91-101.

Blue Chip Economic Indicators, Washington, DC: Capitol Publications, various issues.

Money Rate Report, New Orleans: Commonwealth Research Group, various issues.

Wall Street Journal, 6 July 1987, p. 2.

The Stock Market Premium, Production, and Relative Risk Aversion

By SIMON BENNINGA AND ARIS PROTOPAPADAKIS*

Higher relative risk aversion (RRA) is associated with higher risk premiums only if the riskiness of output is exogenous. When consumers can affect the variability of output, the market risk premium may well decrease as the RRA increases. With constant relative risk aversion and linear production functions, the ratio of the market risk premium to the standard deviation of the market is constant and independent of the RRA. (JEL D20, D80)

Suppose it is possible to observe two economies in equilibrium. In each economy, consumption and production decisions are made by a representative consumer. Suppose that all aspects (production possibilities, initial endowments, etc.) of the two economies are the same, except that the relative risk aversion, henceforth RRA, of the representative consumer in one economy is larger than the RRA of the second economy's representative consumer. In which economy would the equilibrium stock market risk premium be larger?

The intuitive and appealing answer is that the economy with the larger RRA would exhibit the larger market premium. This answer derives in part from the intuition given in papers by David Cass and Joseph E. Stiglitz (1970, 1972) and Peter A. Diamond and Stiglitz (1974), which show

that, given a choice between one risky and one riskless asset, a consumer with higher RRA will choose a less risky portfolio than a similar consumer with a lower RRA. Indeed, most researchers who have tried to measure the average RRA use models whose first-order conditions formalize this intuitive answer. Some examples are Irwin Friend and Marshall E. Blume (1975), Irwin Friend and Joel Hasbrouck (1982), Lars Peter Hansen and Kenneth J. Singleton (1983a,b), Roger-A. Morin and Antonio Fernandez Suarez (1983), Varouj A. Aivazian et al. (1986), Robert H. Litzenberger and Ehud I. Ronn (1986), and George G. Szpiro (1986).

In this paper, we show that the extension of the Cass-Stiglitz results to the general equilibrium pricing of risk can be problematic. Aggregating individual portfolio choices will reveal little about the price of market risk, because a shift in demand from risky to risk-free assets in general will cause the price of the risky asset to rise and that of the risk-free asset to fall. Therefore, the effect on the market risk premium is unclear. However, if the aggregate risk is assumed to be exogenous, then it follows that an economy with a larger RRA will exhibit a larger market risk premium.

While it is natural to assume that the micro risk-return trade-offs are exogenous to individuals' portfolio choices, taking the economy-wide risk-return trade-off as fixed is a much more questionable assumption. For instance, Roger C. Kormendi and Philip

*School of Business, Hebrew University, Jerusalem 91905, Israel, and School of Business, FBE, University of Southern California, Los Angeles, CA 90089. This paper was completed while Benninga was on sabbatical leave at the Anderson Graduate School of Management of the University of California at Los Angeles and Protopapadakis was at the Economics Department, Claremont Graduate School. We acknowledge helpful comments from Michael Brennan, Roger Kormendi, Lillian Ng, Uri Possen, Jeremy Siegel, Dick Startz, Eli Talmor, Mark Weinstein, and seminar participants at the European Finance Association Meetings in Istanbul, the Claremont Graduate School, and the University of Utah. We also thank two anonymous referees for helpful and stimulating comments. The usual disclaimer applies.

G. Meguire (1985) report that the range of output growths observed across countries can be described as choices made from a risk-return frontier: the cost of higher growth is higher variability of output. At the macro level, production is not exogenous, and utility-maximizing economic actors can choose, at least to some extent, the variability of output. When this choice of production variability is taken into account, we demonstrate that it need not follow that higher RRA is associated with a higher risk premium.

In particular, we show that in an economy of a representative consumer with a constant RRA utility function and with linear but risky production possibilities, the equilibrium market risk premium decreases as the RRA increases if the consumer can affect the variability of output. If she cannot, then the market risk premium increases as the RRA increases.

In Section I, we present our basic model. Section II discusses the case of linear production. In Section III, we discuss some implications of the results for the market return:risk ratio; we also discuss how production-function curvature may cancel the effect of utility-function curvature on the riskiness of output.

I. The Model

Throughout the paper we discuss a two-date model.¹ At date 0, the consumer chooses consumption for the current period and inputs which will be used to provide (through a production technology) consumption at date 1. Uncertainty is modeled by assuming that at date 1 one of two states occurs and that the output provided by the production technologies is state-dependent.

We wish to distinguish between *complete* and *incomplete* production technologies. A complete production technology allows the consumer to choose separately the total output in each of the two states occurring at date 1. An incomplete production technol-

ogy does not allow this separation.² Suppose state 1 at date 1 is a "bad" state of the world and state 2 is a "good" state ("good" and "bad" in the sense that the same input will produce more output in the good state than in the bad state). In a complete production technology environment, the consumer can choose the ratio of consumption in the bad state to that in the good state. Thus, the consumer can choose the date-1 variability of her consumption. In an incomplete production environment, the consumer cannot do this: more consumption in the bad state will necessarily mean more consumption in the good state also.

We show that the ability of the consumer to choose consumption variability is a critical factor in determining whether increasing relative risk aversion implies a greater or lesser market risk premium. Although this claim may be illustrated under general conditions, the underlying intuition is most easily grasped in a model with linear production functions. In the next section, we discuss such a model. Before doing so, however, we set out the two cases and derive the first-order conditions.

A. The Complete-Production-Technologies Case

With complete production technologies, the consumer maximizes the utility of consumption today, c_0 , and expected consumption in states 1 and 2 tomorrow, c_1 and c_2 , subject to the following constraints:

$$(1a) \quad \max[EU(c_0, c_1, c_2)] = U(c_0) - \delta\{\pi U(c_1) + (1 - \pi)U(c_2)\}$$

²The standard Arrow-Debreu definition of completeness stresses the ability of consumers to use financial markets to decompose vectors of state-dependent returns into their basis vectors (cf. Gerard Debreu, 1959; Kenneth Arrow, 1964) and to price these basis vectors separately. With a single representative consumer making economy-wide choices, this ability to separate consumption can be achieved only if it exists in the aggregate. Our definition represents an extension of the standard terminology to an economy with production.

¹The results of the paper remain the same in a multiperiod version of the model.

subject to

$$(1b) \quad c_0 = W - z_1 - z_2$$

$$(1c) \quad c_1 = \alpha f(z_1)$$

$$(1d) \quad c_2 = \beta f(z_2).$$

At date 0, the consumer chooses inputs z_1 and z_2 of the model's single physical good. These inputs are used to produce $\alpha f(z_1)$ and $\beta f(z_2)$ at date 1 in states 1 and 2, respectively.³ We assume that the production function is increasing and weakly concave ($f' > 0$, $f'' \leq 0$). We further assume that $\pi\alpha < (1-\pi)\beta$, so that state 1 at date 1 is the bad state and state 2 is the good state. The parameter δ is the consumer's pure time-preference factor. Throughout the paper we look at an isoelastic utility function:

$$(2) \quad U(x) = \frac{x^{1-\gamma}}{1-\gamma} \quad \text{for } \gamma \neq 1$$

$$U(x) = \log(x) \quad \text{for } \gamma = 1.$$

The consumer's relative risk aversion is $-xU''(x)/U'(x) = \gamma$.

The first-order conditions for this problem are

$$(3) \quad q_1 = \frac{\delta\pi U'(c_1)}{U'(c_0)} = \frac{1}{\alpha f'(z_1)}$$

$$q_2 = \frac{\delta(1-\pi)U'(c_2)}{U'(c_0)} = \frac{1}{\beta f'(z_2)}.$$

The marginal rates of substitution (adjusted for probability and time preference) q_1 and

q_2 are the consumer's equilibrium valuations of one unit of the good, contingent on the occurrence of state 1 or state 2, respectively. These state prices may thus be used to price securities with state-contingent payoffs at date 1. It follows that the *risk-free rate* in the economy is $r_f \equiv 1/(q_1 + q_2) - 1$. The *present value of the market's output* is $PV \equiv q_1\alpha f(z_1) + q_2\beta f(z_2)$. The *expected value of the market's future output* is $EV \equiv \pi\alpha f(z_1) + (1-\pi)\beta f(z_2)$. The *expected return on the market* is $ER_M \equiv EV/(z_1 + z_2) - 1$, and the *market risk premium* is $RP \equiv ER_M - r_f$.

B. The Incomplete-Production-Technologies Case

We contrast the simple case given above with the case in which the consumer has the choice about how much to invest but cannot affect the variability of her future consumption.⁴ We model this case by assuming that there is a single production technology which produces, with different efficiencies, in each of the two states. Formally, the consumer maximizes

$$(4a) \quad \max[EU(c_0, c_1, c_2)] = U(c_0) + \delta\{\pi U(c_1) + (1-\pi)U(c_2)\}$$

subject to

$$(4b) \quad c_0 = W - z$$

$$(4c) \quad c_1 = \alpha f(z)$$

$$(4d) \quad c_2 = \beta f(z)$$

where, as before, $f(z)$ is weakly concave and increasing.

³The assumption that none of the z_i is productive in state j ($i, j = 1, 2$) is an inessential simplifying assumption; what is crucial to our results is that the productivity of z_i is less than z_j in state j . The results hold for more complex linear production technologies, which produce in both states of the world, as long as these technologies span the desired consumption point, thus admitting an unconstrained (non-corner-solution) optimum.

⁴For example, in the macroeconomic literature, researchers typically introduce uncertainty by specifying a production function of the form $y_t = f(k_t) + \epsilon_t$. This specification is equivalent to assuming that consumers cannot affect the riskiness of output; this holds even if the equations are in logs, in which case the error term multiplies the quantities in levels.

The first-order condition for this problem is

$$(5) \quad \frac{\delta\{\pi\alpha U'(c_1) + (1-\pi)\beta U'(c_2)\}}{U'(c_0)} = \frac{1}{f'(z)}.$$

As before, the state prices q_1 and q_2 are the probability and time-discount adjusted marginal rates of substitution, $\delta\pi U'(c_1)/U'(c_0)$ and $\delta(1-\pi)U'(c_2)/U'(c_0)$, respectively. The expected return on the market for this case is $ER_M = [\pi\alpha f(z) + (1-\pi)\beta f(z)]/z - 1$. The definitions for the riskless rate of interest and the market risk premium remain unchanged.

II. The Case of Linear Production Functions

In this section, we consider the two models set out above for the special case when $f(z)$ is linear and passes through the origin [i.e., $f(z) = z$].

A. The Complete-Markets Case

The first model we consider is one in which production technologies are complete. Thus, the representative consumer can choose the ratio of state 1 to state 2 consumption at the second date. Rewriting the first-order conditions gives

$$(6a) \quad (\delta\pi\alpha)^{1/\gamma}(W - z_1 - z_2) = \alpha z_1$$

$$(6b) \quad [\delta(1-\pi)\beta]^{1/\gamma}(W - z_1 - z_2) = \beta z_2.$$

Letting $X = (\delta\pi\alpha)^{1/\gamma}$ and $Y = [\delta(1-\pi)\beta]^{1/\gamma}$, the solution to these equations is given by

$$(7) \quad z_1 = \frac{\beta X W}{\alpha\beta + X\beta + \alpha Y}$$

$$z_2 = \frac{\alpha Y W}{\alpha\beta + X\beta + \alpha Y}.$$

THEOREM 1: *When the production function is linear and passes through the origin, the market risk premium declines as γ increases.*

PROOF:

It follows from the assumptions of the theorem that the risk-free rate is independent of the inputs and is given by $r_f = 1/[1/\alpha + 1/\beta] - 1$. One plus the expected return of the market is given by

$$(8) \quad 1 + ER_M = \frac{\pi\alpha z_1 + (1-\pi)\beta z_2}{z_1 + z_2}$$

$$= \frac{\alpha\beta[\pi X + (1-\pi)Y]}{X\beta + \alpha Y}.$$

Differentiating equation (8) with respect to γ (ignoring the $\alpha\beta$ terms and the denominator of the derivative) gives that $dRP/d\gamma$ is proportional to

$$(9) \quad [\alpha\pi - (1-\pi)\beta] \left[-1/\gamma^2 \right]$$

$$\times XY \{ \ln[\delta\pi\alpha] - \ln[\delta(1-\pi)\beta] \}$$

which is always negative.

The intuition behind Theorem 1 is the following: as the RRA is raised, the consumer wishes to make the variance of the consumption smaller. She does this by increasing the ratio z_1/z_2 . This has the effect of shifting resources to the relatively inefficient production process, and it reduces the average return on investment and with it the total market return. Since the risk-free rate in the complete-markets model is independent of the consumer's maximization (due to the linear technology), it follows that the market risk premium must fall. Another way to see this is to notice that the consumer can vary the ratio of the maximum quantity of the risk-free asset to her total wealth. The maximum quantity of the risk-free asset is $\alpha z_1/(1+r_f)$, since $\beta z_2 > \alpha z_1$. As she becomes more risk-averse, the consumer demands (and gets) more of the risk-free asset at the cost of a lower average return on her investment.⁵

⁵A useful decomposition of the representative consumer's portfolio is that it contains a risk-free asset with payoffs αz_1 in both states, whose value is $\alpha z_1/(1+r_f)$, and also contains an option whose payoff

$$(13) \quad \frac{d[q_1 + q_2]}{d\gamma} = \frac{[\pi\alpha^{1-\gamma} + (1-\pi)\beta^{1-\gamma}][-\pi\alpha^{-\gamma}\ln(\alpha) - (1-\pi)\beta^{-\gamma}\ln(\beta)]}{[\pi\alpha^{1-\gamma} + (1-\pi)\beta^{1-\gamma}]^2} \\ - \frac{[\pi\alpha^{-\gamma} + (1-\pi)\beta^{-\gamma}][-\pi\alpha^{1-\gamma}\ln(\alpha) - (1-\pi)\beta^{1-\gamma}\ln(\beta)]}{[\pi\alpha^{1-\gamma} + (1-\pi)\beta^{1-\gamma}]^2}$$

In a partial-equilibrium framework, Cass and Stiglitz (1970, 1972), Diamond and Stiglitz (1974), and Friend and Blume (1975) show that, when an individual consumer's RRA increases, the proportion of the risky asset held in the consumer's portfolio will decrease. The implications of this result for the market risk premium are unclear, since the shift in asset demand will tend to raise the price of the risk-free asset and lower the price of the risky asset. Theorem 1 shows that, if the economy can be modeled by a representative consumer and complete production technologies, the risk premium declines.

B. The Incomplete-Markets Case

In the incomplete-markets case with a linear production function, the input of z at date 0 will produce output of αz in state 1 and output of βz in state 2 of the next period. There is no possibility for the consumer to alter the ratio of consumption in the two states. One plus the expected return on the market is given by

$$(10) \quad 1 + ER_M = \frac{\pi c_1 + (1-\pi)c_2}{z} \\ = \pi\alpha + (1-\pi)\beta.$$

To derive the riskless rate for this case, note that the first-order condition for z simplifies to give

$$(11) \quad \delta\{\pi\alpha^{1-\gamma}z^{-\gamma} + (1-\pi)\beta^{1-\gamma}z^{-\gamma}\} = (W-z)^{-\gamma}.$$

is 0 in state 1 and $\beta z_2 - \alpha z_1$ in state 2, whose value is $(\beta z_2 - \alpha z_1)/\beta$. An increase in the representative consumer's RRA has the effect of lowering both the option's payoff and its value.

Using equation (11), we may now write

$$(12) \quad q_1 + q_2 = \frac{\delta\pi[\alpha z]^{-\gamma}}{[W-z]^{-\gamma}} + \frac{\delta(1-\pi)[\beta z]^{-\gamma}}{[W-z]^{-\gamma}} \\ = \frac{\pi\alpha^{-\gamma} + (1-\pi)\beta^{-\gamma}}{\pi\alpha^{1-\gamma} + (1-\pi)\beta^{1-\gamma}}.$$

The theorem below follows from equation (12).

THEOREM 2: *When production technologies are incomplete, the market risk premium increases as the RRA increases.*

PROOF:

In this case, the expected market return is invariant to the risk aversion. Taking the derivative of (12) gives equation (13), above. The denominator of this expression is positive, and the numerator simplifies to $\pi(1-\pi)\alpha^{-\gamma}\beta^{-\gamma}[\alpha-\beta][\ln(\alpha)-\ln(\beta)]$. This expression is positive, since $\alpha-\beta$ and $\ln(\alpha)-\ln(\beta)$ have the same sign. It thus follows that the numerator increases with γ , which shows that r_f decreases with increasing risk aversion. Since the expected return on the market is fixed [equation (10)], this proves the result.

As the RRA of the representative consumer increases, she will be unable to influence the ratio c_2/c_1 of consumption at the next date. This is the essence of "incompleteness" of production technologies. In order for the markets to clear, the consumer must be satisfied with this risk configuration. This can only happen if the risk premium rises enough to compensate her for the undesired risk. Since the total market return is fixed, this change can be accomplished only with a lower risk-free rate. Another way to see why the risk-free rate

TABLE 1—PRODUCTION TECHNOLOGIES

RRA	Complete markets				Incomplete markets			
	r_f	ER_M	σ	RP	r_f	ER_M	σ	RP
1	0.008	0.8297	1.2262	0.8217	0.0588	0.0658	0.0862	0.0070
2	0.008	0.3581	0.5224	0.3501	0.0519	0.0658	0.0862	0.0139
3	0.008	0.2246	0.3232	0.2166	0.0452	0.0658	0.0862	0.0206
4	0.008	0.1640	0.2328	0.1560	0.0388	0.0658	0.0862	0.0270
5	0.008	0.1297	0.1817	0.1217	0.0326	0.0658	0.0862	0.0332
6	0.008	0.1077	0.1488	0.0997	0.0269	0.0658	0.0862	0.0389
7	0.008	0.0924	0.1260	0.0844	0.0215	0.0658	0.0862	0.0443
8	0.008	0.0812	0.1092	0.0732	0.0166	0.0658	0.0862	0.0492
9	0.008	0.0726	0.0964	0.0646	0.0121	0.0658	0.0862	0.0537
10	0.008	0.0658	0.0862	0.0578	0.0080	0.0658	0.0862	0.0578
15	0.008	0.0459	0.0565	0.0379	-0.0065	0.0658	0.0862	0.0723
20	0.008	0.0361	0.0420	0.0281	-0.0140	0.0658	0.0862	0.0798
25	0.008	0.0304	0.0334	0.0224	-0.0175	0.0658	0.0862	0.0833
30	0.008	0.0266	0.0277	0.0186	-0.0191	0.0658	0.0862	0.0849
35	0.008	0.0239	0.0237	0.0159	-0.0199	0.0658	0.0862	0.0857
40	0.008	0.0219	0.0207	0.0139	-0.0202	0.0658	0.0862	0.0860
45	0.008	0.0203	0.0184	0.0123	-0.0203	0.0658	0.0862	0.0861
50	0.008	0.0191	0.0165	0.0111	-0.0204	0.0658	0.0862	0.0862

Notes: This table compares equilibrium values when α and β are held constant and RRA is allowed to vary in the complete- and incomplete-markets cases. In the complete-markets case $\alpha = 1.2071$ and $\beta = 6.1118$; in the incomplete-markets case $\alpha = 0.9796$ and $\beta = 1.1520$. These values have been chosen because in both cases they produce an expected market return of 6.58 percent and a risk-free rate of 0.8 percent for RRA = 10, when $\delta = 0.95$. The risk-free rate is invariant to RRA in the complete-markets case, and the expected return on the market is invariant in the incomplete-markets case. The table entries for σ are the standard deviations of the market returns for each case; in the incomplete-markets case, σ is given by the standard deviation of α and β . As shown in the paper, the ratio RP/σ is a constant in the complete-markets case; this is not always obvious in the table, because of rounding error.

falls is to notice that, unlike in the case with complete production markets, the ratio of the maximum quantity of the risk-free asset to total wealth is fixed by technology. If the consumer becomes more risk-averse, the equilibrium price of the risk-free asset must increase, and its return must decrease. Theorem 2 parallels partial-equilibrium results by David P. Baron (1970) for individual firms.⁶

C. Simulating a Linear System

In Table 1, we give some sample simulations for complete- and incomplete-technologies cases. In both systems, α and β have

been chosen so that, with a relative risk aversion of 10 and $\delta = 0.95$, the equilibrium will produce a real risk-free rate of 0.8 percent and an expected return on the market of 6.58 percent.⁷ In the complete markets case, α and β are the simultaneous solution of $1 + r_f = 1/[1/\alpha + 1/\beta] - 1$ and equation (8), for the target values of ER_M and r_f . In the incomplete-markets case, α and β solve equations (10) and (12) simultaneously. For the target values of $ER_M = 6.58$ percent and $r_f = 0.8$ percent, this gives $\alpha = 1.2071$ and $\beta = 6.1118$ in the complete-markets case and $\alpha = 0.9796$ and $\beta = 1.1520$ in the incomplete-markets case. The values of α and β differ for the two cases because in the case of complete production technologies the consumer can achieve the target market return by an allocation of her

⁶Another related result is that of Robert B. Barsky (1989), who shows that in an endowment economy very similar to our incomplete-markets economy, a mean-preserving spread in the risky asset's returns will depress the risk-free rate, although its effect on the market risk premium is ambiguous.

⁷These values are the historic averages in the United States for the period 1898–1978. For details, see Rajnish Mehra and Edward C. Prescott (1985).

production inputs that is skewed toward the low-productivity state. In the case of incomplete production technologies, the achievement of the target is only possible if production efficiency is much more nearly equal in both states.

As proved above, the risk premium is declining in RRA in the complete-markets case and increasing in RRA in the incomplete-markets case.⁸

III. Some Implications of Our Results

A. The Market Return: Risk Ratio

Endowment models of asset pricing typically yield relations in which the market risk premium is proportional to market variance and RRA. For example, Robert Merton (1973) shows that if the risky asset's returns are lognormally distributed and the consumer maximizes an isoelastic utility function, then $(ER_M - r_f)/\sigma_M^2 = RRA$. In the Merton framework, higher RRA results in a higher risk premium, holding variance constant. When the production technologies are endogenized, as in our model with complete production technologies, an entirely different result holds. Under these conditions, the ratio of the market risk premium to the market standard deviation is independent of the relative risk aversion. We prove this result in the following theorem.

THEOREM 3: *In a complete-production-technologies equilibrium with linear production functions, the ratio of the market risk premium to the standard deviation of the market returns is constant.*

PROOF:

For the linear complete-markets case,

$$(14) \quad 1 + r_f = \frac{\alpha\beta}{\alpha + \beta}.$$

⁸The results in Table 1 are to be interpreted as an illustration of the effect of risk aversion on the market risk premium for complete and incomplete markets; for similar calculations see Miles S. Kimball (1988), Benninga and Protopapadakis (1990), and other references cited in the Introduction of this paper.

It follows from equations (8) and (14) that the market risk premium is:

$$(15) \quad RP = \frac{\alpha\beta[\pi X + (1-\pi)Y]}{X\beta + \alpha Y} - \frac{\alpha\beta}{\alpha + \beta} \\ = \frac{\alpha\beta[(1-\pi)\beta - \pi\alpha](Y - X)}{(X\beta + \alpha Y)(\alpha + \beta)}.$$

To compute the standard deviation of the market return, note that the expressions for 1 plus the state-1 and state-2 returns are

$$(16) \quad \frac{\alpha z_1}{z_1 + z_2} = \frac{\alpha\beta X}{\alpha X + \beta Y} \\ \frac{\beta z_2}{z_1 + z_2} = \frac{\alpha\beta Y}{\alpha X + \beta Y}.$$

The standard deviation of the market return, σ_M , can be computed from equation (16):

$$(17) \quad \sigma_M = \frac{\alpha\beta(Y - X)[\pi(1-\pi)]^{1/2}}{\alpha X + \beta Y}.$$

Thus, RP/σ_M is given by

$$(18) \quad \frac{RP}{\sigma_M} = \frac{[(1-\pi)\beta - \pi\alpha]}{(\alpha + \beta)[\pi(1-\pi)]^{1/2}}.$$

This proves the theorem.

Theorem 3 is consistent with one of the results reported by Kenneth R. French et al. (1987). They report better fits when they regress the market risk premium on the standard error rather than on the variance of market returns (see also Merton, 1980; Lillian Ng, 1988). Theorem 3 suggests that a linear regression of the risk premium on the market variance may not be well specified, because the ratio of the risk premium to the variance is itself a function of the RRA, unlike the ratio of the risk premium to the standard error.

B. Nonlinear Production Functions

In Section II, we showed that the risk premium is independent of wealth and falls as the RRA rises, when the production technologies are complete and linear. Since linear technologies imply that the risk-free rate is independent of input choices, when the expected market return falls so does the risk premium. However, when production functions are concave, the risk-free rate depends on production choices. Hence, a change in input choice that leads to lower variability of output will lead to a lower return on the market and also to a lower risk-free rate. If the production technologies are sufficiently concave, the risk-free rate may fall faster than the total market return, so that the risk premium can rise. Since the equilibrium in general will depend on the wealth of the consumer, it follows that the relation between the risk premium and the RRA depends on wealth for complete, concave production technologies.⁹

Some researchers have tested the specification of an isoelastic utility function by testing whether wealth enters significantly in asset pricing equations. This may not be a useful specification test, since the risk premium should be a function of wealth even with isoelastic utility functions, provided there are nonlinearities in production.

IV. Conclusion

General equilibrium macroeconomic models with uncertainty are often cast in the form of microeconomic paradigms of consumer behavior toward risk. We show that this procedure implies some very strong assumptions about the economic environment and that it can lead to misleading conclusions. It is natural to model portfolio choices at the micro level as purely risk-

sharing behavior where risk and return are exogenous to the individual decision-makers. In contrast, macro behavior under uncertainty is primarily about the endogenous choice of aggregate risk and return. In an important sense, macro equilibrium models are about risk-choosing behavior.

In this paper, we demonstrate the importance of this distinction. When the representative consumer is allowed to choose the variability of output, the market risk premium generally declines as the risk aversion of the consumer increases. This happens because a more risk-averse representative consumer will choose lower variability of aggregate output at the expense of lower expected return. However, when the consumer's choice is limited to the pricing of exogenous risk, the market risk premium increases as the risk aversion of the consumer increases, because a more risk-averse consumer requires a higher premium in order to hold willingly the fixed amount of risk.

Finally, the model has interesting implications for the testing of alternative hypotheses about the market return:risk ratio and the relation between the market risk premium and wealth.

REFERENCES

- Aivazian, Varouj A., Callen, Jeffrey L., Krinsky, Itzhak and Kwan, Clarence C. V., "An Empirical Portfolio Analysis of Financial Asset Substitutability: The Case of the U.S. Household Sector," *Quarterly Review of Economics and Business*, Summer 1986, 26, 47-65.
- Arrow, Kenneth J., "The Role of Securities in the Optimal Allocation of Risk-Bearing," *Review of Economic Studies*, April 1964, 31, 91-6.
- Baron, David P., "Price Uncertainty, Utility, and Industry Equilibrium in Pure Competition," *International Economic Review*, October 1970, 11, 463-80.
- Barsky, Robert B., "Why Don't the Prices of Stocks and Bonds Move Together?" *American Economic Review*, December 1989, 79, 1132-45.

⁹If the production functions "flatten out" as the inputs to the functions get larger (i.e., $f''' > 0$), it may be shown that the market risk premium will be a decreasing function of RRA for large-enough initial consumer wealth.

- Benninga, Simon and Protopapadakis, Aris, "Leverage, Time Preference, and the 'Equity Premium Puzzle,'" *Journal of Monetary Economics*, January 1990, 25, 49-58.
- Cass, David and Stiglitz, Joseph E., "The Structure of Investor Preferences, Asset Returns, and Separability in Portfolio Allocation," *Journal of Economic Theory*, June 1970, 2, 122-60.
- _____ and _____, "Risk Aversion and Wealth Effects on Portfolios with Many Assets," *Review of Economic Studies*, July 1972, 39, 331-54.
- Debreu, Gerard, *Theory of Value*, New York: Wiley, 1959.
- Diamond, Peter A. and Stiglitz, Joseph E., "Increases in Risk and in Risk Aversion," *Journal of Economic Theory*, July 1974, 8, 337-60.
- French, Kenneth R., Schwert, G. William and Stambaugh, Robert F., "Expected Stock Returns and Volatility," *Journal of Financial Economics*, September 1987, 19, 3-30.
- Friend, Irwin and Blume, Marshall E., "The Demand for Risky Assets," *American Economic Review*, December 1975, 65, 900-22.
- _____ and Hasbrouck, Joel, "The Effect of Inflation on the Profitability and Valuation of U.S. Corporations," in Marshall Sarnat and Giorgio Szego, eds. *Savings, Investment, and Capital Markets in an Inflationary Economy*, Cambridge: Ballinger, 1982, pp. 37-119.
- Hansen, Lars Peter and Singleton, Kenneth J., (1983a) "Generalized Instrument Variables Estimation of Nonlinear Expectations Models," *Econometrica*, September 1983, 50, 1269-86.
- _____ and _____, (1983b) "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns," *Journal of Political Economy*, April 1983, 91, 249-65.
- Kimball, Miles S., "Precautionary Saving and the Marginal Propensity to Consume," mimeo, University of Michigan, 1988.
- Kormendi, Roger C. and Meguire, Philip G., "Macroeconomic Determinants of Growth: Cross-Country Evidence," *Journal of Monetary Economics*, September 1985, 16, 141-60.
- Litzenberger, Robert H. and Ronn, Ehud I., "A Utility-Based Model of Common Stock Price Movements," *Journal of Finance*, March 1986, 41, 67-92.
- Mehra, Rajnish and Prescott, Edward C., "The Equity Premium: A Puzzle," *Journal of Monetary Economics*, March 1985, 15, 145-61.
- Merton, Robert C., "An Intertemporal Capital Asset Pricing Model," *Econometrica*, September 1973, 41, 867-87.
- _____, "On Estimating the Expected Return on the Market," *Journal of Financial Economics*, December 1980, 8, 323-61.
- Morin, Roger-A. and Fernandez Suarez, Antonio, "Risk Aversion Revisited," *Journal of Finance*, September 1983, 38, 1201-16.
- Ng, Lillian, "Expected Risk Premium and Implied Market Volatility," unpublished manuscript, The Wharton School of the University of Pennsylvania, 1988.
- Szpiro, George G., "Measuring Risk Aversion: An Alternative Approach," *Review of Economics and Statistics*, February 1986, 68, 156-9.

The Temporal Stability of Dividends and Stock Prices: Evidence from the Likelihood Function

By DAVID N. DEJONG AND CHARLES H. WHITEMAN*

The debate over whether the expected present value of dividends adequately describes stock prices hinges in part on whether dividends are trend-stationary or integrated processes: it does not if dividends are trend-stationary; it does if they are integrated. This paper argues that classical statistical tests only indicate that there is not sufficient evidence to reject either specification and provides Bayesian analyses designed to reveal the relative support the data give to the two specifications. The analysis suggests that dividends and prices are more likely to be trend-stationary than integrated, leaving the determination of prices a puzzle. (JEL E30, C12, C22)

The temporal stability of dividends and stock prices is an issue central to the recent debate over the validity of the perfect-markets hypotheses (PMH). This hypothesis, sometimes called the efficient- or rational-markets hypothesis, is that stock market returns are not forecastable. If the hypothesis holds, stock prices should equal the present value of expected future dividends. Yet in a widely cited paper in this *Review*, Robert J. Shiller (1981a) assumed that dividends are *trend-stationary* (stationary about a deterministic trend) and found prices to be too volatile to be consistent with the hypothesis. In contrast, the results of Terry A. Marsh and Robert C. Merton (1986) in a subsequent *AER* paper exactly reversed Shiller's, provided that dividends feed back on prices, and are thus *integrated*.

This controversy over the PMH has been intensified by three recent extensions of the

temporal-stability issue. First, the empirical validity of the trend-stationarity assumption has been challenged: dividends and stock prices have been subjected to and have generally passed ever more sophisticated tests for integration. For example, Allan W. Kleidon (1986) drew the inference of integration and concluded that Shiller's results are accounted for by the unit root rather than violations of the PMH. Second, the importance of the trend-stationarity assumption has been downplayed: tests of the PMH which *condition* on integration have been developed. However, while Kleidon's (1986) conditional variance-inequality tests (valid under integration) did not find evidence against the PMH, the cointegration test of John Y. Campbell and Shiller (1987a) rejected restrictions the PMH places on the temporal behavior of prices and dividends when both are integrated. Third, implications of the PMH which hold for both trend-stationary and integrated dividend processes have been investigated.

The import of these extensions hinges on the quality of the evidence concerning the integration of prices and dividends: that of the first two clearly so, and that of the third to the extent that the temporal stability issue remains unresolved. Given the stature of the PMH in economics and finance, it is thus not surprising that enormous efforts have been expended to develop and to apply tests for integration. Substantial progress

*Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260; and Department of Economics, University of Iowa Iowa City, IA 52242. Support from the National Science Foundation under grants SES 85-10505 and 89-22419 to The University of Iowa and 90-05180 to The University of Pittsburgh is gratefully acknowledged. We thank Forrest Nelson, Gene Savin, and anonymous referees for helpful comments and Robert Shiller for providing us with his Dow Jones and Standard and Poor's data sets. Our approach to this problem owes much to the encouragement of John Geweke.

has been made on the asymptotic-distribution theory for ordinary least-squares regression in the presence of unit roots (e.g., Peter C. B. Phillips, 1987, 1988), yet applications have been less than satisfactory. Conflicting evidence is prevalent in the literature, and there seems to be disagreement over what inferences should be drawn.

We study the temporal stability of stock prices and dividends using an alternative procedure based on the likelihood principle; our inferences are *conditional*, given observed data. That is, we ask what the data tell us about the process most likely to have generated the data we observe. In fact, we follow Christopher A. Sims (1988) and Sims and Harald Uhlig (1988) in taking an explicitly Bayesian approach: we summarize sample information in a likelihood function which we feel most researchers would be willing to consider, and we specify prior distributions over exponential and stochastic trends in the time series. The prior distributions we specify render posterior analysis analytically impossible; hence, we follow Teun Kloek and Herman K. van Dijk (1978) and John Geweke (1986, 1988, 1989a,b) in performing necessary integrations by Monte Carlo methods.

Figures 1C–6C depict posterior densities for the dominant autoregressive root (λ ; less than unity if the series is trend-stationary) and the coefficient on the exponential trend (β_4) for several aggregate price and dividend series. The prior used to construct the distributions places substantial weight on the unit root but is uninformative concerning the trend on the indicated domain. The inference we draw is that the series are not integrated: for practically *any* prior that assigns some weight to a positive deterministic trend, the data drive the posterior for the dominant root away from unity and drive the posterior for the deterministic trend away from zero. A very sharp prior on trendless random walks will support the posterior inference of a unit root; this prior is in fact the null hypothesis used in classical tests of the unit-root hypothesis. However, since trend-stationarity is the alternative of interest, with such a sharp prior, why look at the data at all?

Classical procedures have often failed to find evidence against unit roots; this has led some researchers to infer that the stock market data are in fact integrated. Yet similar classical procedures also fail to find evidence against the trend-stationarity assumption (see DeJong et al., 1989a). Aside from inducing the standard tongue-twisting classical result that failure to reject each of two mutually exclusive hypotheses does not settle the issue between them, this pair of findings underscores the importance of considering the *relative* support the competing specifications receive from available evidence. Our Bayesian procedures are designed to do exactly this, and they clearly suggest that, relative to the integration specification, the trend-stationary specification is more likely.

1. Integration, Trend-Stationarity, and the PMH

The debate over the PMH comprises perhaps the best-developed sequence of argument-counterargument in recent empirical economics. To understand the nature of the debate and the contribution of this paper, it will prove useful to review the literature briefly but with some precision.

In support of his assumption of trend-stationarity, Shiller (1981b) tested and rejected the hypothesis of integration for Standard and Poor's (S&P's) dividends over the years 1872–1978. Specifically, using the model

$$(1) \ln(D_t) = \mu + \gamma t + \rho \ln(D_{t-1}) + \varepsilon_t$$

he rejected the integration hypothesis ($\rho = 1$, $\gamma = 0$) using “unit-root” tests developed by Wayne A. Fuller (1976) and David A. Dickey and Fuller (1981); he then proceeded in his (1981a) analysis under the assumption that Dow Jones and S&P's prices and dividends are trend-stationary.

Kleidon (1986) argued that Shiller's (1981a) data are in fact integrated. Using a shorter S&P's data set (1926–1979), Kleidon estimated

$$(2) \ln(X_t) = \delta + \alpha \ln(X_{t-1}) + \xi_t$$

for both prices and dividends and failed to

reject the integration hypothesis ($\alpha = 1$). Further, Kleidon (1986 p. 993) argued that rejections based on (1) are not relevant, since including time as a regressor adds virtually no explanatory power over model (2) and since results derived from (2) are "economically more reasonable" than the results derived from (1).

Kenneth West (1987) and Pierre Perron (1988) argued that unit-root tests based on (2) are inappropriate and misleading when the alternative is trend-stationarity. In short, if $\ln(X_t)$ is trend-stationary, equation (2) is a misspecification: the only way to fit growth in the series in this case is for estimates of α to be driven erroneously to unity, which causes tests based on (2) to be inconsistent. The remedy is to use (1). Perron did so for the S&P's 1871–1984 data using Phillips and Perron's (1988) generalizations of the Dickey and Fuller (1981) tests appropriate when ε_t is serially correlated; he rejected the null for dividends.

Campbell and Shiller (1987b) provided evidence similar to Perron's for the S&P's 1871–1986 data and also tested for integration in dividends and prices using the value-weighted and equally-weighted New York Stock Exchange (NYSE) 1926–1986 data sets. Using the S&P's data, they failed to reject the null for prices but rejected it for dividends; using the NYSE data, they failed to reject the null for both prices and dividends. Alternatively, Campbell and Shiller (1987a) tested for integration in the *levels* of S&P's 1871–1986 prices and dividends. This time they *failed* to reject the null for either prices *or* for dividends.

DeJong et al. (1989a,b) showed that the integration tests employed in these studies, while consistent because the "extraneous" trend is included, can have lower power against plausible trend-stationary alternatives, making nonrejections suspect. DeJong et al. (1989a) also developed a relatively more powerful classical test of the hypothesis $\gamma \neq 0$, $\rho = 0.85$ in (1), which permits serial correlation. They failed to reject the null for the price and dividend series in the Dow, NYSE, or S&P's data sets.

The findings reported above are summarized in Table 1. Results are mixed: there

are three rejections and eight nonrejections of the integration hypothesis, and there are six nonrejections of the trend-stationarity hypothesis evaluated by DeJong et al. (1989a). Inferences concerning integration are evidently fragile, depending upon details of the data used, the time horizon considered, corrections for nuisance parameters, and, we think, the strong priors of researchers. Thus, the status of the debate over Shiller's (1981a) study and Marsh and Merton's (1986) response remains in doubt.

The other extensions have not settled the issue. Kleidon (1986) failed to find evidence against the PMH in the short S&P's data using a conditional volatility test, while Campbell and Shiller (1987a) did find such evidence in the long S&P's data using a cointegration test. Neither approach is valid if the data are trend-stationary. West (1988) tested a volatility relation that is valid under either trend-stationarity or integration and rejected the relation for the 1871–1980 S&P's and the Dow data sets. Yet while the relation West tested is robust to the specification of dividends, the procedure he employed is not: if the dividends West differenced are actually trend-stationary, important parameters used in the construction of the test are estimated inconsistently.¹

Finally, Eugene F. Fama and Kenneth R. French (1988), Andrew W. Lo and A. Craig MacKinlay (1988), and James M. Poterba and Lawrence H. Summers (1988) have developed "variance-ratio" tests which suggest that stock prices exhibit mean-reverting behavior. The evidence in these studies suggests that weekly and monthly returns do revert; hence, the PMH appears suspect. In

¹In a related context, Marjorie A. Flavin (1983) has argued that Shiller's (1981a,b) procedures are plagued by sampling difficulties; thus, even data that satisfy the PMH could appear excessively volatile, though apparently not by the magnitudes detected by Shiller. N. Gregory Mankiw et al. (1985) conducted a test that is, in effect, a hybrid of the West (1988) test and Kleidon's (1986) conditional-variance test. However, they did not report any information concerning the sampling distribution of their test statistic; hence, their results are difficult to interpret.

TABLE 1—EXISTING INTEGRATION AND TREND-STATIONARITY RESULTS

Data	Test	Result	Reference
Standard and Poor's dividends (1871–1978)	equation (1); $H_0: \rho = 1$	H_0 rejected at 5-percent level	Shiller (1981b)
Standard and Poor's dividends and prices (1926–1979)	equation (2); $H_0: \alpha = 1$	cannot reject H_0 at 10-percent level for dividends or prices	Kleidon (1986)
Standard and Poor's dividends and prices (1871–1984)	equation (1); $H_0: \rho = 1$	H_0 rejected at 5-percent level for dividends; not rejected for prices at 10-percent level	Perron (1988)
Standard and Poor's dividends and prices (1871–1986)	equations (1) and (2); H_0 : random walk	cannot reject H_0 at 10-percent level for dividends or prices	Campbell and Shiller (1987a)
Standard and Poor's dividends and prices; levels (1871–1986)	equation (1); $H_0: \gamma = 0, \rho = 1$	H_0 rejected at 5-percent level for dividends; not rejected for prices at 10-percent level	Campbell and Shiller (1987b)
NYSE value- and equally-weighted dividends and prices (1926–1986)	equation (1); $H_0: \gamma = 0, \rho = 1$	cannot reject H_0 at 10-percent level for p or d	Campbell and Shiller (1987b)
Dow, NYSE, and S&P's dividends and prices ¹ (see Appendix for dates)	equation (1); $H_0: \gamma \neq 0, \rho = 0.85$	cannot reject H_0 at 5-percent level for p or d	DeJong et al. (1989a)
Monthly NYSE stock returns (1926–1985)	H_0 : no mean reversion	H_0 rejected	Fama and French (1988)
Weekly NYSE stock returns (1962–1985)	H_0 : no mean reversion	H_0 rejected	Lo and MacKinlay (1988)
Monthly NYSE stock returns (1871–1986)	H_0 : no mean reversion	H_0 rejected	Poterba and Summers (1988)
Annual Dow, NYSE, and S&P's returns (see Appendix for dates)	H_0 : no mean reversion	H_0 not rejected	DeJong and Whiteman (1990)

contrast, DeJong and Whiteman (1990) found no mean reversion in the annual Dow, NYSE, and S&P's price and dividend data sets. Yet so long as the temporal-stability issue remains unresolved, the mean-reversion studies constitute *the* evidence against the PMH.² We now turn to an attempt to settle the issue.

II. Bayesian Inference in a Normal Time-Series Model³

According to the classical metric of sampling distributions, integrated series and trend-stationary series are not dissimilar, and the stock market data are consistent with either hypothesis. Of course, we only

²In a recent paper in this *Review*, Stephen G. Cecchetti et al. (1990) showed that the degree of mean reversion that characterizes stock returns could be well within the margins of sampling error.

³This section makes use of standard Bayesian results (e.g., Arnold Zellner, 1971; Edward E. Leamer, 1978) and also draws heavily on Geweke (1986, 1988, 1989a, b).

have one sample, and given the importance of the temporal-stability issue, it is necessary to draw an inference one way or the other. Classical sampling-theory procedures do not allow such an inference; thus, it is useful to explore other methods of assessing the relative likelihood of the competing specifications.

In specifying the likelihood function that will form the basis for our analysis, we follow common practice and adopt an autoregressive (AR) specification. Denote by y_t the natural logarithm of real dividends (or stock prices). The evolution of y_t is assumed to be described by the autoregression

$$(3) \quad y_t = \beta_0 + \beta(L)y_t + \beta_4 t + \varepsilon_t$$

$$\varepsilon_t \sim \text{i.i.d. } \mathcal{N}(0, \sigma^2)$$

where $\beta(L) = \beta_1 L + \beta_2 L^2 + \beta_3 L^3$; y_0, y_{-1} , and y_{-2} are fixed; and the lag operator L is defined by $L^n y_t = y_{t-n}$.⁴ Three lags were chosen to allow for serial correlation in y_t . Stacking the T observations in the standard fashion, we have $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.⁵ We define

$$(4) \quad \lambda(z) \equiv 1 - \beta(z) = \prod_{j=1}^3 (1 - \lambda_j z).$$

The roots of the AR representation of $\{y_t\}$ correspond to the p factors of $\lambda(z)$ or, equivalently, the roots of $F^p \lambda(F)$, where $F = z^{-1}$. We denote these roots by $\lambda_1, \lambda_2, \dots, \lambda_p$ and define the dominant autoregressive root by $\Lambda = \max_j |\lambda_j|$. Trend-stationarity corresponds to $\Lambda < 1$; integration (or difference-stationarity) is the special case $\Lambda = 1$. A formulation that makes

trend-stationarity the special case is considered in Section IV.

Sample information is summarized in the likelihood function

$$\ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = (2\pi\sigma^2)^{-T/2} \times \exp\left[-(1/2\sigma^2)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right].$$

The classicist views the likelihood function as the sampling distribution for the data $\{\mathbf{y}, \mathbf{X}\}$ given a fixed value of $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}', \sigma^2)$. Alternatively, the Bayesian views the likelihood function as an ingredient in the computation of the conditional probability density for $\boldsymbol{\theta}$ given $\{\mathbf{y}, \mathbf{X}\}$. In particular, with prior views about $\boldsymbol{\theta}$ summarized in the density $p(\boldsymbol{\theta})$, the posterior density for $\boldsymbol{\theta}$ is

$$P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})p(\boldsymbol{\theta})/f(\mathbf{y}, \mathbf{X})$$

$$\propto \ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})p(\boldsymbol{\theta})$$

where $f(\mathbf{y}, \mathbf{X})$ is a constant from the point of view of the $\boldsymbol{\theta}$ distribution.

The prior distribution we posit is the typical noninformative one: flat over the β_j and $\ln(\sigma)$ (i.e., $p(\boldsymbol{\theta}) = \sigma^{-1}$). Conditioned on σ , the posterior distribution for $\boldsymbol{\beta}$ is normal with mean \mathbf{b} and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The marginal density for σ , $P(\sigma|\mathbf{y}, \mathbf{X})$, is inverted gamma: $\sigma^2 = \nu s^2/u$, where u is distributed as χ^2 with ν degrees of freedom.

In practice, the prior we employed was given by $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta})\iota_A(\boldsymbol{\theta})$, where $\iota_A(\boldsymbol{\theta})$ is the indicator function for the event A . For example, one such A is given by

$$A = \{(\beta_4, \Lambda): 0 \leq \beta_4 \leq 0.016, 0.55 \leq \Lambda \leq 1.055\}.$$

If $g(\boldsymbol{\theta})$ is any function of interest [e.g., $g(\boldsymbol{\theta}) = (\Lambda, \beta_4)$], then

$$E[g(\boldsymbol{\theta})|A, \mathbf{y}, \mathbf{X}]$$

$$= \frac{\int_{\boldsymbol{\theta} \in A} g(\boldsymbol{\theta}) \ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta} \in A} \ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

⁴A referee points out that the homoscedasticity assumption concerning ε_t in (3) might usefully be relaxed. Doing so would forfeit comparability with the classical unit-root tests discussed above and lies beyond the scope of this paper. Further, Geweke (1989a) seems to indicate that relatively little is gained (for IBM prices) by admitting ARCH. Still, such an extension is part of planned future work.

⁵Specifically, $\mathbf{y} = (y_1 \dots y_T)'$, $\mathbf{X} = (\mathbf{x}_1' \dots \mathbf{x}_T')'$ with $\mathbf{x}_t = (1 \ y_{t-1} \ y_{t-2} \ y_{t-3} \ t)'$, and $\boldsymbol{\beta} = (\beta_0 \ \beta_1 \ \beta_2 \ \beta_3 \ \beta_4)'$.

TABLE 2—ROOT CALCULATIONS, TREND EXCLUDED

Series	μ_Λ	σ_Λ	$p(\Lambda \geq 0.975 \beta_4 = 0)$	$p(0.65 \leq \Lambda < 1.16 \beta_4 = 0)$
Dow Jones dividends	0.916	0.071	0.190	0.992
Dow Jones prices	0.898	0.082	0.159	0.980
NYSE dividends	0.984	0.031	0.645	1.000
NYSE prices	0.971	0.037	0.473	1.000
Standard and Poor's dividends	0.953	0.032	0.235	1.000
Standard and Poor's prices	0.951	0.031	0.217	1.000

Notes: μ_Λ denotes the posterior mean of the dominant root computed from the AR(3) representation of the series (with trend excluded), and σ_Λ denotes the posterior variance. These results are based on 20,000 replications.

TABLE 3—ROOT CALCULATIONS, TREND INCLUDED

Series	μ_Λ	σ_Λ	μ_{β_4}	σ_{β_4}	ρ_{Λ, β_4}	$p(\Lambda \geq 0.975 A)$	$p(\beta_4 \leq 0.001 A)$	$p(A)$
Dow Jones dividends	0.717	0.096	0.006	0.003	-0.504	0.004	0.018	0.872
Dow Jones prices	0.764	0.105	0.006	0.003	-0.531	0.011	0.045	0.873
NYSE dividends	0.768	0.103	0.008	0.003	-0.753	0.010	0.012	0.921
NYSE prices	0.835	0.097	0.007	0.003	-0.745	0.032	0.033	0.909
Standard and Poor's dividends	0.719	0.091	0.003	0.001	-0.657	0.000	0.023	0.911
Standard and Poor's prices	0.869	0.065	0.002	0.001	-0.686	0.015	0.241	0.961

Notes: μ_Λ and σ_Λ are as described in Table 2, and μ_{β_4} and σ_{β_4} are analogous for the posterior exponential trend of the series; ρ_{Λ, β_4} is the posterior correlation between the trend coefficient and dominant root. These results are based on 20,000 replications. The event A is $\{0 \leq \beta_4 \leq 0.016, 0.55 \leq \Lambda \leq 1.055\}$.

and

$$PR[\theta \in A | y, X] = \frac{\int_{\theta \in A} \ell(\theta | y, X) p(\theta) d\theta}{\int_{\theta} \ell(\theta | y, X) d\theta}.$$

The integrals cannot be evaluated analytically, but integration by Monte Carlo (Kloek and van Dijk, 1978; Geweke, 1986, 1988, 1989a, b) is straightforward. That is, given a sequence of independent drawings of random variables from the posterior distribution $P(\theta | y, X)$, we can estimate these integrals by averaging across drawings.⁶ Geweke

(1989b) shows that these sample averages are consistent estimators of the integrals themselves (in Monte Carlo replications). Moreover, in computing the probabilities of various events, we follow Geweke (1989b) and note that each drawing can be considered a Bernoulli trial with success probability q . The natural estimate of q is the number of successes divided by the number of trials, and the standard error of this estimate is $[q(1-q)/n]^{1/2}$. Thus, the standard error of the estimate of q after $n = 10,000$ is no greater than 0.005.

In the next section, we present numerical estimates of the joint densities, $P(\Lambda, \beta_4 | y, X)$, and use them to draw inferences concerning difference- versus trend-stationarity. We infer that the integration representation is not implausible if there is

⁶Details of this procedure are spelled out in Geweke (1989b) and DeJong and Whiteman (1991).

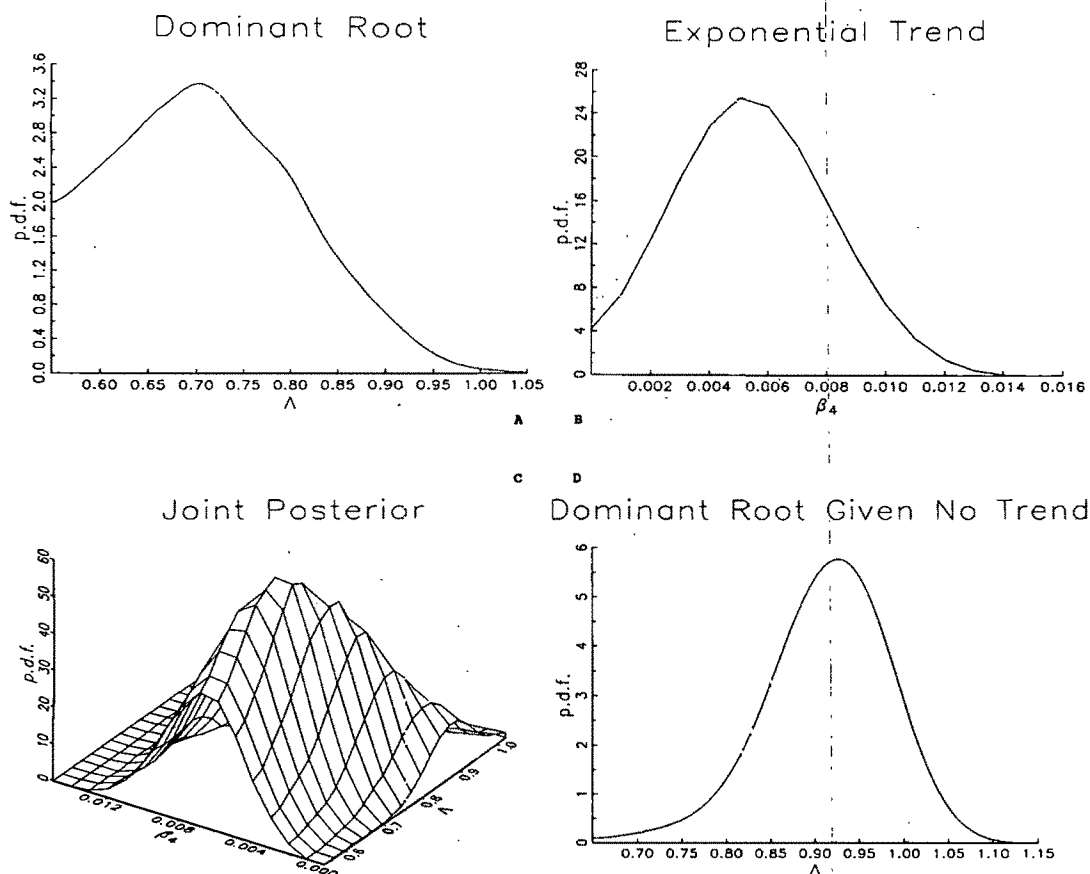


FIGURE 1. DOW JONES DIVIDENDS 1928-1973

at least a 5-percent posterior probability that Λ exceeds 0.975.

III. Inference Concerning Exponential Trend and the Dominant Root

We conduct posterior inference for aggregate annual dividends and stock prices from the Dow, the NYSE, and S&P's indexes (the data are described in the Appendix). We consider two priors; we first restrict $\beta_4 = 0$ and then view β_4 as uniform on $[0, 0.015]$. The associated posterior distributions for the exponential trend (β_4) and dominant AR root (Λ) are summarized in Tables 2 and 3 and in Figures 1-6.⁷

⁷The posteriors were calculated as follows. A 101×16 grid was created over $\{(\Lambda, \beta_4) \in [0.55, 1.05] \times$

In each figure, panel A presents the marginal posterior probability density function (p.d.f.) for Λ , panel B presents the marginal density for β_4 , panel C presents the joint (Λ, β_4) posterior, and panel D presents the density for Λ obtained using the zero-trend prior. The densities represent the combination of the raw histograms with a locally quadratic smoothness prior.⁸

$[0, 0.015])$. For each of 20,000 replications, (Λ, β_4) was calculated, and posterior frequencies were accumulated in the 1,616 bins (of width 0.005 in the Λ -dimension and 0.001 in the β_4 -dimension).

⁸The smoothness priors are akin to those used by Shiller (1973) in estimating distributed lags, and they have a "mixed estimation" interpretation (Henri Theil and A. H. Goldberger, 1961). Let \mathbf{h} be a 101×1 vector containing raw histogram values, and let \mathbf{D} be a $101 \times$

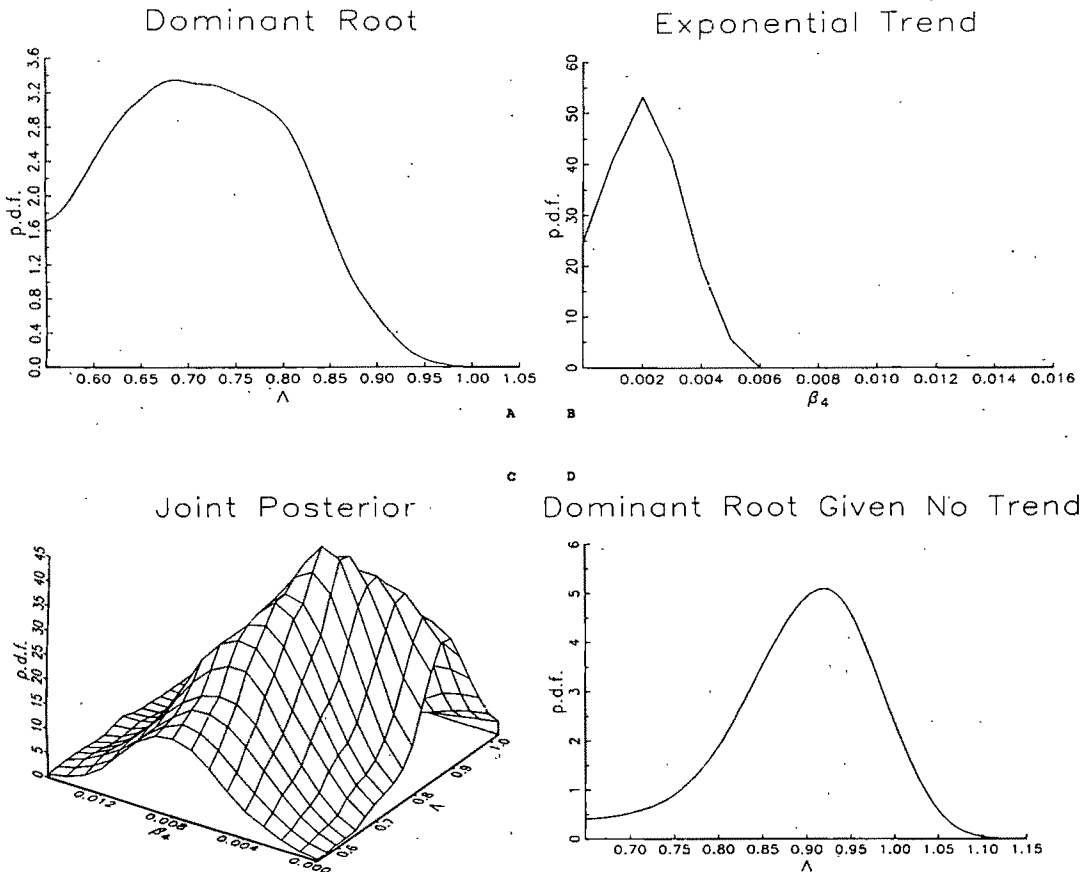


FIGURE 2. DOW JONES PRICES 1928-1978

The first finding of interest is that the price and dividend series do not give rise to very different posteriors. With either of the two trend priors, the price and dividend posteriors are similarly shaped and located.

The posteriors of Λ obtained under the zero-trend prior strongly support integration; the posterior means range from 0.898

(for Dow dividends) to 0.971 (for NYSE prices) and are each within 1.65 standard deviations of unity. Further, the posterior probability that Λ exceeds 0.975 is no less than 15 percent. This result corresponds to Kleidon's (1986) finding that, when the deterministic trend alternative is excluded from the analysis, the integration representation is supported.

When we relax the zero-trend prior, the posterior modes of the exponential trends (Figs. 1B-6B) always exceed zero. For the S&P series, the trend is small, but the posterior density suggests a value of 0.001 or 0.002. Under this prior, the integration hypothesis is not supported; the posterior means for Λ range from 0.717 (standard deviation = 0.096) in the Dow dividends se-

101 matrix with first row $(-1 \ 1 \ 0 \ \dots \ 0)$, second row $(1 \ -2 \ 1 \ 0 \ \dots \ 0)$, third row $(0 \ 1 \ -2 \ 1 \ 0 \ \dots \ 0)$, ..., 100th row $(0 \ \dots \ 0 \ 1 \ -2 \ 1)$, and last row $(0 \ \dots \ 0 \ 1 \ -1)$. Note that Dh is a vector of second differences of the histogram values (end points excepted). The prior is that $Dh \sim \mathcal{N}(0, \kappa^2)$. Let $Z = [ID'/\kappa]$. The smoothed density is given by $h^* = (Z'Z)^{-1}Z'h$. The figures use $\kappa = 10$.

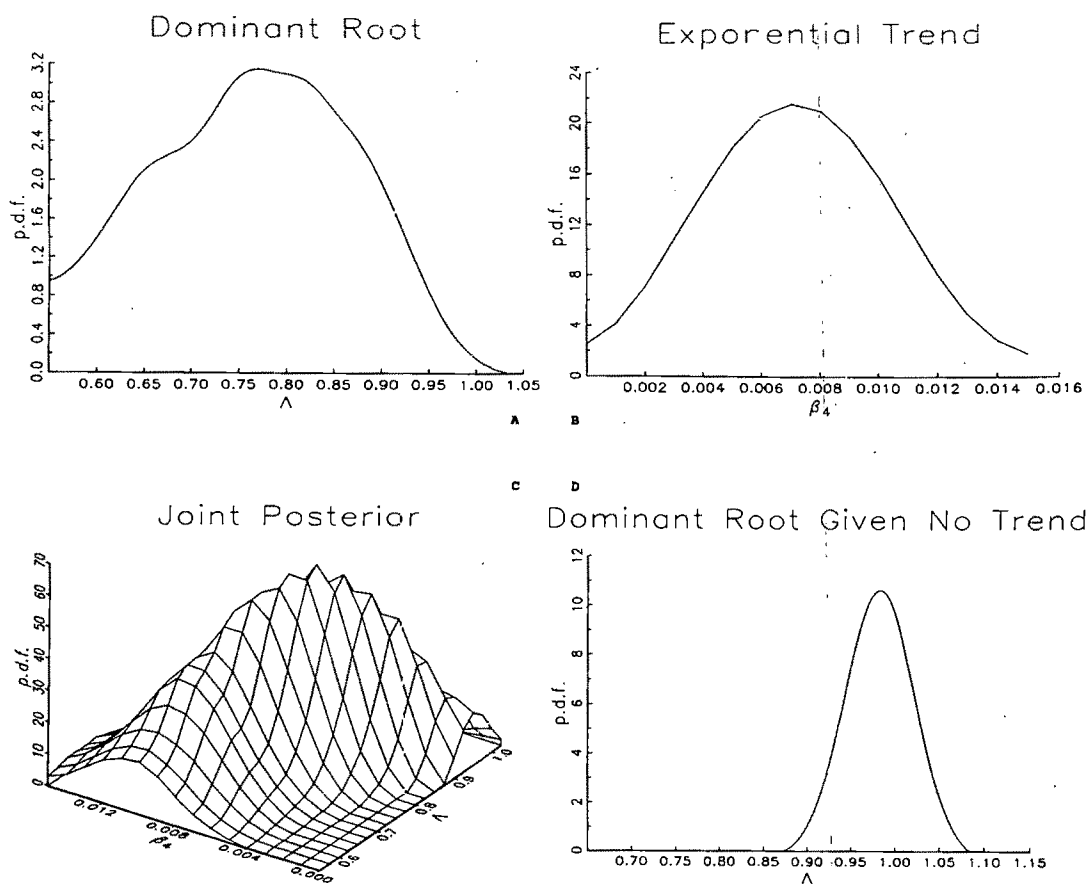


FIGURE 3. NYSE DIVIDENDS 1926-1981

ries to 0.869 (SD = 0.065) in the S&P price series. Further, the root distributions are skewed left; hence, the upper 1.65σ bound never exceeds unity. Moreover, given the restriction to Λ (minimum probability 0.87), the estimated probability that Λ is greater than 0.975 never exceeds 3.2 percent.

One of the most striking results is the strong negative correlation between Λ and β_4 . This correlation ranges from -0.504 (Dow dividends) to -0.753 (NYSE dividends), and it is clear that restricting the exponential trend to zero (an inference wildly at variance with the data, having posterior probability greater than 0.033 in only one case [S&P prices], when it reaches 0.24) produces estimates of Λ that are far too

large. Hence, while in Table 2 and Figures 1D-6D it appears that unit roots pervade the stock price and dividend series, this is an artifact of the unsupported restriction $\beta_4 = 0$; when the trend is unrestricted, the dominant roots appear to be about 0.8.

The figures suggest the following inference: with no apparent particular initial feeling concerning the trend and dominant root, one would be moved by the data toward the trend-stationary specification. Yet while the prior we consider is flat over the parameters β , and in particular the trend coefficient β_4 , it is *not* flat over Λ , which represents a nonlinear transformation of these parameters. Figure 7A illustrates the shape of our prior over Λ , which was calcu-

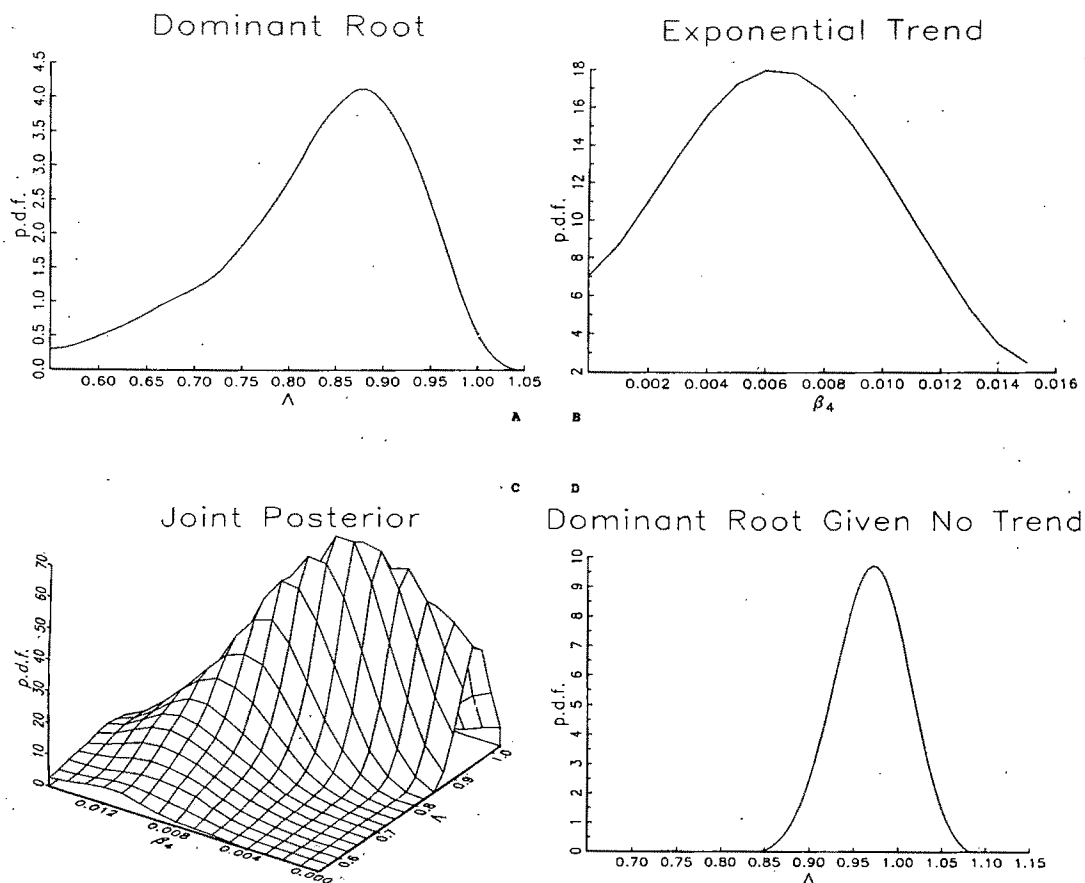


FIGURE 4. NYSE PRICES 1926-1981

lated by the Monte Carlo method described above, except that the drawings of β were obtained from a uniform distribution.⁹ Note that increasingly large values of Λ receive increasing prior weight. This is a "Marsh-Merton" prior: integration is relatively supported over trend-stationarity. Despite this support, the posteriors we obtain clearly favor the inference of trend-stationarity. Recalculating the posteriors under a flat prior over dominant roots would involve

downweighting large roots and increasing weights on small roots; the posteriors would shift leftward, and unit roots would look even less plausible.

IV. Extensions

In any statistical analysis, results can be sensitive to (i) the researcher's prior views and (ii) specification of the likelihood function. To investigate the robustness of our results to the prior, we consider the *coherence* of our procedure: what prior ensures against the trend-stationarity inference in repeated samples from difference-stationary processes? To examine the robustness to the specification of the likelihood function,

⁹The reason we calculate our "prior" over Λ is that it is a very complicated nonlinear transformation of β : roots of a polynomial are calculated, an absolute value is taken, and to top things off, a maximum is taken.

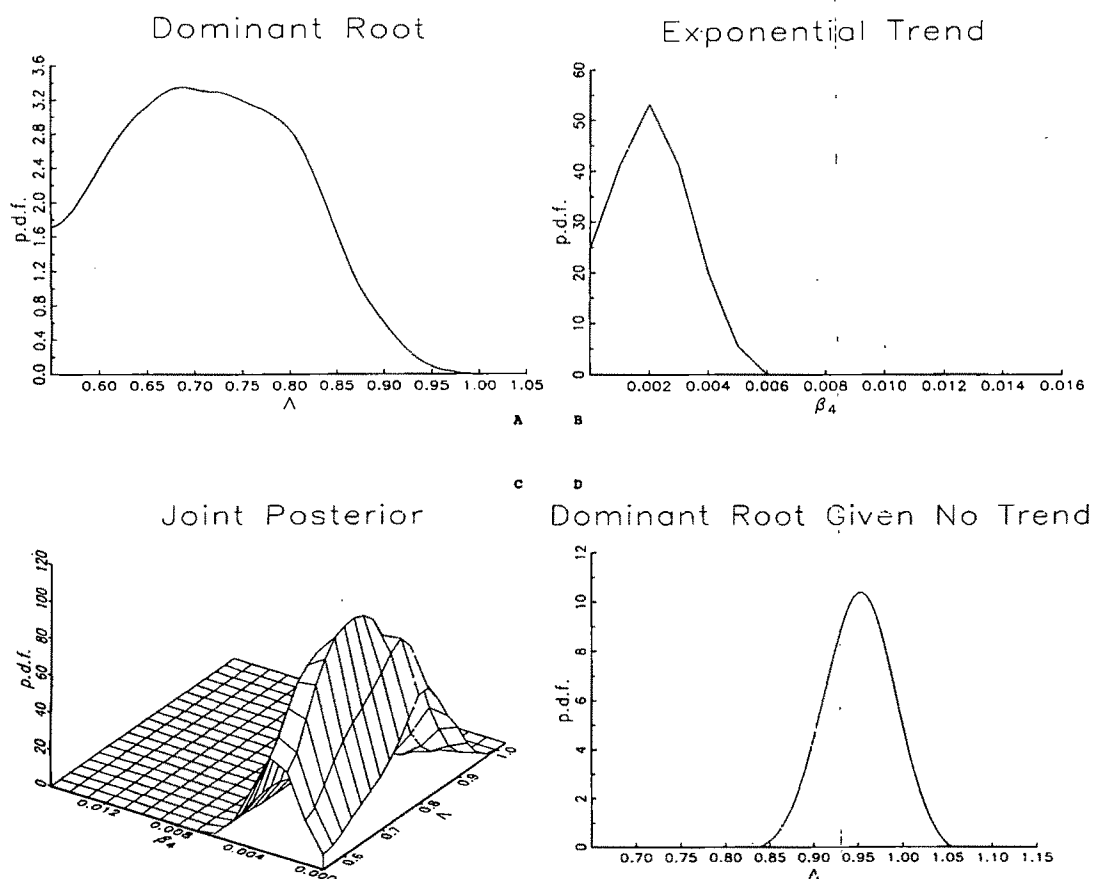


FIGURE 5. STANDARD AND POOR'S DIVIDENDS 1871-1985

we consider an unobserved-stochastic-trend generalization of (3).

A. Robustness to Prior

The coherence issue was investigated following DeJong and Whiteman (1991), in which we developed a prior distribution over (Λ, β_4) which insures that our inferential procedure errs no more than 5 percent of the time in repeated samples when the series under consideration is of length 100 and follows a random walk with drift and unit innovation variance. The prior over Λ is the same as that pictured in Figure 7A, but assigns decreasing weight to increasingly large trend coefficients. Specifically, the prior over β_4 is a beta distribution propor-

tional to

$$(1 - \beta_4)^{(B-1)} \quad \beta_4 \in [0, 0.015]$$

with $B = 6$; this "5-percent prior" is illustrated in Figures 7B-D. Reweighting our original posteriors using the 5-percent prior, we find that the integration representation continues to appear implausible for each series except NYSE prices; the posterior probability that Λ exceeds 0.975 is 6.1 percent in this case (see Table 4 for details).

B. Robustness to Specification

While the approaches taken by classical unit-root testers do suggest substantial agreement over the form of the likelihood

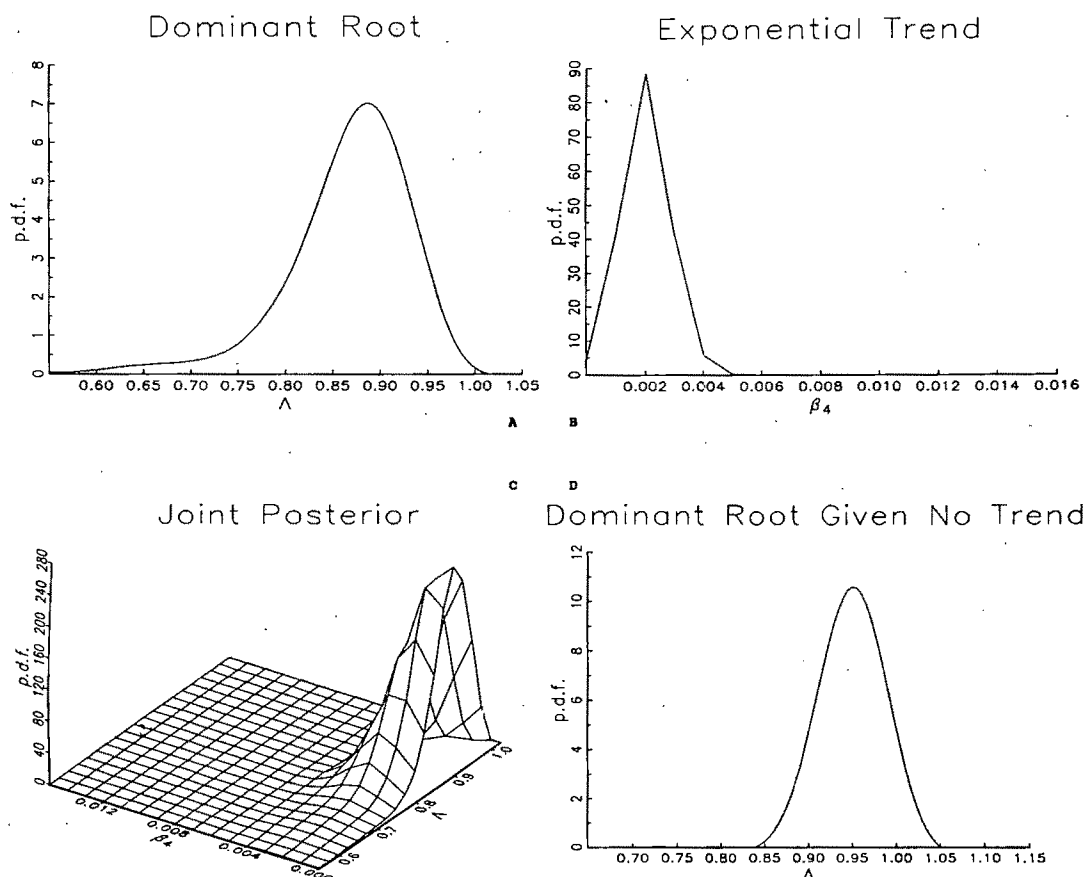


FIGURE 6. STANDARD AND POOR'S PRICES 1871-1985

function, the unit root remains a special case in these (and our) formulations. Thus, to investigate the sensitivity of our results to the likelihood function utilized above, we consider a generalization that makes trend-stationarity the special case. In particular, we consider an unobserved components model for y_t of the form

$$(5) \quad y_t = \gamma + \mu_t + \Psi_t$$

where

$$(6) \quad \mu_t = \delta + \mu_{t-1} + \eta_t$$

$$\eta_t \sim \text{i.i.d.} \mathcal{N}(0, \sigma_\eta^2)$$

$$(7) \quad \Psi_t = \phi_1 \Psi_{t-1} + \phi_2 \Psi_{t-2} + \phi_3 \Psi_{t-3} + \varepsilon_t$$

$$\varepsilon_t \sim \text{i.i.d.} \mathcal{N}(0, \sigma_\varepsilon^2)$$

and $\{\eta_t\}$ and $\{\varepsilon_t\}$ are independent. Notice that if $\sigma_\eta^2 = 0$, then $\mu_t = \mu_0 + \delta t$, and y_t has the AR(3) representation given in (3). When $\sigma_\eta^2 > 0$, y_t possesses a stochastic trend and can be represented as ARIMA(3,1,3):

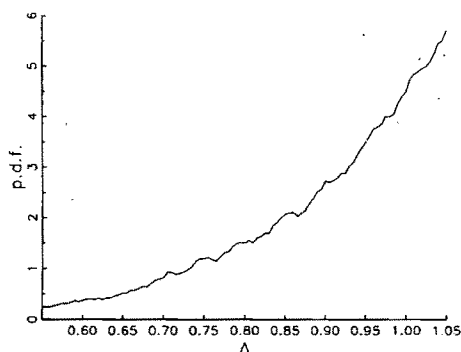
$$(8) \quad \Delta y_t = \delta + \phi(L)^{-1} \theta(L) \varepsilon_t$$

where $\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \theta_3 L^3$, $\{\varepsilon_t\}$ is the one-step-ahead forecast error in predicting $\{y_t\}$ linearly from its own past, and using the notational convention $|f(z)|^2 = f(z)f(z^{-1})$,

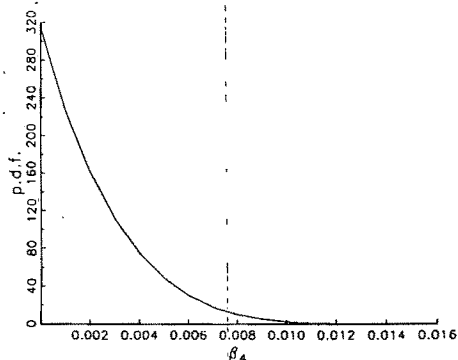
$$(9) \quad |\theta(z)|^2 \sigma_\varepsilon^2 = |\phi(z)|^2 \sigma_\eta^2 + |1 - z|^2 \sigma_\varepsilon^2.$$

We denote the dominant moving-average root by $\lambda = 1/\max_z |\arg[\theta(z) = 0]|$.

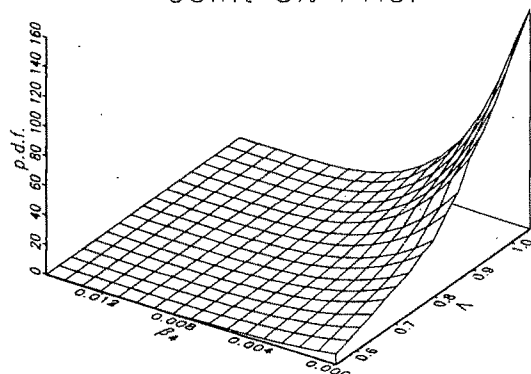
Prior Distribution in Root Space



Prior Distribution in Trend Space



Joint 5% Prior



Equal-Probability Countours

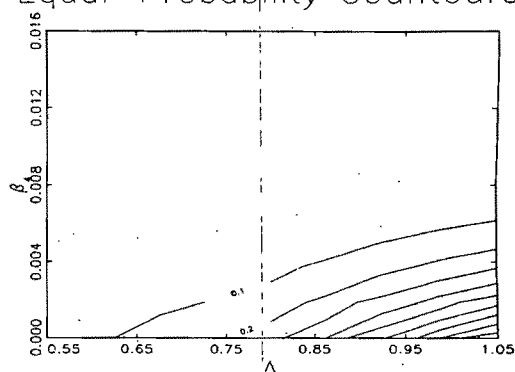


FIGURE 7. PRIOR DISTRIBUTIONS

Expression (9) indicates the factorization problem that must be solved to determine $\theta_1, \theta_2, \theta_3$, and σ_e^2 from the parameters $\phi_1, \phi_2, \phi_3, \sigma_\eta^2$, and σ_ε^2 of (5). Note that when $\sigma_\eta = 0$, the MA representation for Δy_t has a zero at unity. Thus, there are *two* senses in which trend-stationarity is the special case in this formulation: $\sigma_\eta = 0$ and $\lambda = 1$.

Define $\Theta = (\phi_1 \phi_2 \phi_3 \sigma_\eta \sigma_\varepsilon)$, stack the deviations from the mean of the T observations $\Delta y_1, \dots, \Delta y_T$ in the $T \times 1$ vector y , and write $Eyy' = \sigma_e^2 V$. The likelihood function of y under this specification is given by

$$\log \mathcal{L}(\Theta | y) = -(T/2) \log 2\pi - (T/2) \log \sigma_e^2 - \frac{1}{2} \log |V| - \frac{1}{2} \sigma_e^2 y' V^{-1} y.$$

Posterior analysis was conducted by combining $\mathcal{L}(y | \Theta)$ with an uninformative prior over Θ .¹⁰

We focused on three parameters in our analysis: the dominant AR root λ , the dominant MA root λ , and the ratio of the standard deviation of the stochastic-trend innovation to the total standard deviation of the series, $\sigma_\eta / \sigma_\varepsilon$. The prior and posterior distributions for these parameters were ob-

¹⁰The prior was

$$\pi(\beta) \propto \begin{cases} \sigma_e^{-1} & \forall \Theta \text{ such that } \lambda \in [0.5, 1], \\ & \lambda \in [0.55, 1.05], \sigma_\eta \geq 0, \sigma_\varepsilon \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

TABLE 4—EXTENSIONS

Series	$p(\Lambda > 0.975 $ 5-percent prior)	Unobserved-components model					
		$\sigma_\eta / \sigma_\epsilon$		λ		Λ	
		Mean	Mode	Mean	Mode	Mean	$p(\Lambda \geq 0.975)$
Dow Jones dividends	0.033	0.226 (0.155)	0.06	0.952 (0.046)	1.00	0.814 (0.099)	0.038
Dow Jones prices	0.045	0.321 (0.197)	0.19	0.958 (0.044)	1.00	0.887 (0.097)	0.179*
NYSE dividends	0.040	0.222 (0.152)	0.07	0.955 (0.044)	1.00	0.789 (0.103)	0.023
NYSE prices	0.061*	0.261 (0.175)	0.12	0.963 (0.039)	1.00	0.870 (0.088)	0.095*
Standard and Poor's dividends	0.002	0.167 (0.113)	0.02	0.965 (0.027)	1.00	0.735 (0.090)	0.001
Standard and Poor's prices	0.021	0.190 (0.127)	0.07	0.979 (0.018)	1.00	0.890 (0.053)	0.039

Notes: The 5-percent prior ensures that our decision rule errs no more than 5 percent of the time in repeated samples from a random walk with drift and unit innovation variance. The decision rule was: draw the unit-root inference if there is at least 5-percent posterior probability on dominant roots exceeding 0.975. Asterisks denote situations in which the unit-root inference is *not* implausible. For the unobserved-components model, $\sigma_\eta / \sigma_\epsilon$ is the ratio of standard deviations of the stochastic-trend innovation to the total innovation in the series estimated from (5), λ denotes the dominant root of the MA component estimated in the model, and Λ denotes the dominant root of the AR component estimated in the model. Posterior standard deviations are given in parentheses.

tained using the numerical integration methods described above.¹¹ A complication in generating posteriors in this case is that it is not possible to generate drawings of Θ directly from the posterior distribution $p(\Theta|y)$; this problem was sidestepped using the importance-sampling techniques described by Geweke (1989b) (see DeJong and Whiteman [1989] for details).

The results obtained using the unobserved-components model are summarized in Table 4 and Figure 8. Consider first the posterior distributions of Λ . While the prior over Λ is the same as in the AR(3) case, the posteriors are rightshifted relative to the AR(3) distributions. In particular, the posterior probability that Λ exceeds 0.975 is greater than 5 percent for two series under this specification: Dow prices (0.179) and NYSE prices (0.095). Under our inferential procedure, we conclude that the integration

representation does not appear implausible for these series in this case; however, the posteriors continue to suggest that the *most likely* values for Λ are less than unity. The integration representation continues to appear implausible for the remaining series.

Consider now the results obtained for $\sigma_\eta / \sigma_\epsilon$. While the prior provides considerable support for the significance of the stochastic trend (the prior mode of $\sigma_\eta / \sigma_\epsilon$ is approximately 0.4), the posteriors provide much weaker support. In fact, the posterior mode of $\sigma_\eta / \sigma_\epsilon$ exceeds 0.1 in only the Dow and NYSE price series, where the posterior modes are 0.19 and 0.12. Moreover, the posterior distributions of λ have a mode of unity for each series and are concentrated much more tightly near unity than is the prior over λ . Since the MA representation of Δy_t has a zero at unity when σ_η is small, this suggests that if there are stochastic-trend components in these series, they are small.

To summarize, using the AR(3) likelihood function, integration looked quite implausible; with the unobserved-components ARIMA(3, 1, 3) specification, integration continues to look implausible, but perhaps

¹¹The unobserved components formulation adds to the complexity of the prior; again, we computed the implications that flat priors over the ϕ 's carry for parameters of interest, λ and Λ .

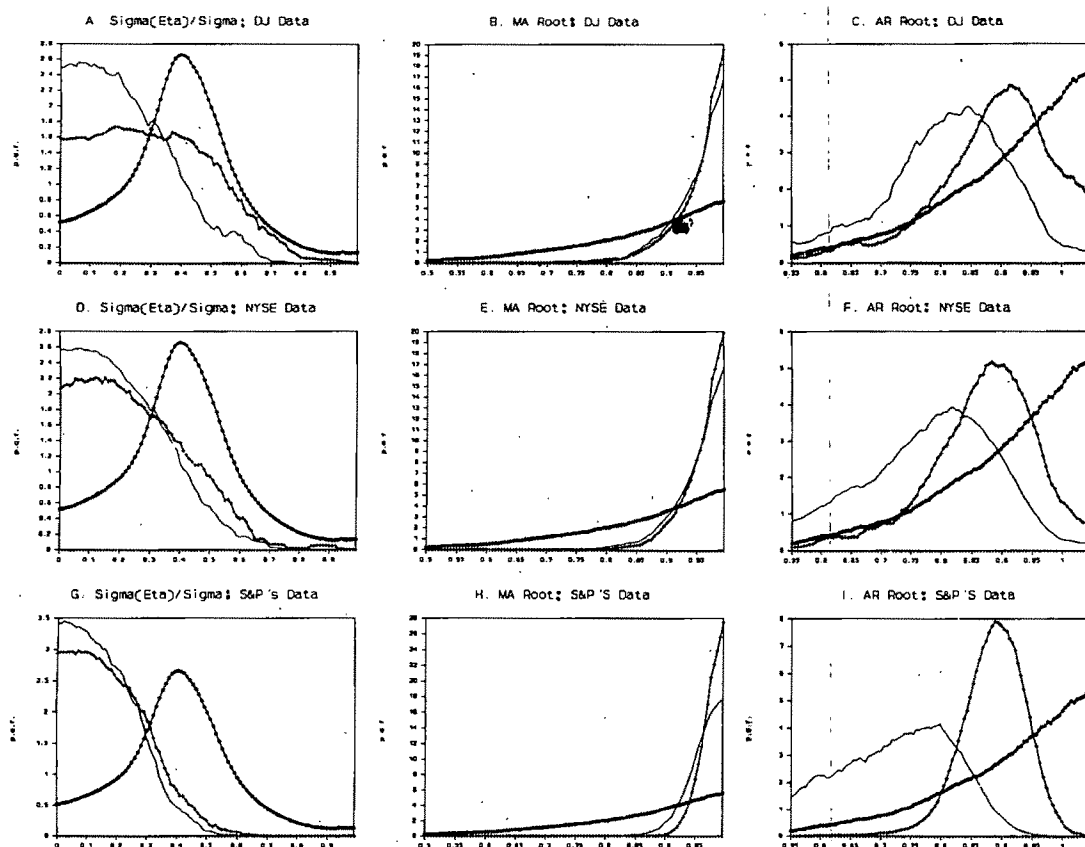


FIGURE 8. UNOBSERVED-COMPONENTS MODEL: — = DIVIDENDS; + = PRICES; ◇ = PRIOR

less strikingly so. That there is some sensitivity to specification is a result not unlike those of John H. Cochrane (1988) and Lawrence J. Christiano and Martin Eichenbaum (1989). Yet the unobserved-components priors we used were quite sharp in favor of integration: the dominant AR root prior peaks at unity, and the prior mode for the stochastic trend innovation is distinctly positive. Despite these prior views in favor of integration, the data drive one away from this, toward trend-stationarity. *Very* strong prior views in favor of integration may not be shaken, but Figure 8 suggests that moderately strong ones would be.

V. Concluding Remarks

The results in this paper seem to encompass existing integration results. When the

exponential trend is restricted to zero, our Figures 1D–6D and Table 2 suggest that the dividend and stock price series are integrated. This corresponds to Kleidon's (1986) results (and his "prior" concerning the presence of trends). But on either classical or Bayesian grounds, the zero-trend restriction is erroneous: Classical unit-root tests which exclude deterministic trends are inconsistent against trend-stationary alternatives, and our posterior distributions suggest that the zero-trend restriction is inappropriate. When the trend coefficient is not restricted, classical unit-root tests do not reject the null (but have low power), classical trend-stationarity tests do not reject the null (and have moderate power), and our posteriors suggest that trend-stationarity is strongly supported. Moreover, upon differencing the series, we find evidence of unit

dominant moving-average roots, suggesting that differencing one time is too many.

This of course is the real issue: how does one handle trends in the data? The series may grow because they are deterministically trended or because they are drifting random walks. Using a classical unit-root test, one examines whether there is evidence *against* the drifting random walk. If one takes the opposing view, adopting trend-stationarity as the null hypothesis, classical tests examine whether there is evidence *against* trend-stationarity. The classical story ends here: neither test rejects, and hence one is left with insufficient evidence to discard either of two mutually exclusive hypotheses. To settle the issue, it is necessary to know which of the two competing specifications is more likely. The Bayesian procedures we adopted enable us to discover this; for the real stock price and dividend series, the trend-stationary specification seems more plausible than the integration specification. Thus, the appropriate inference, we think, is that deterministic exponential trends pervade the data, and dominant roots are less than unity. Thus, Shiller's (1981a) exponential detrending of the data seems justified, and the perfect-markets puzzle remains, unexplained by the integration of dividends.

APPENDIX

A. Data Set 1: Modified Dow Jones Industrial Average (Annual 1928–1978)

Here, P_t and D_t refer to the real price and dividends of the portfolio of 30 stocks comprising the sample for the Dow Jones Industrial Average when it was created in 1928. However, due to the fact that adjustments are continually made to the stocks used to calculate this average, Shiller has extensively modified it to control for these adjustments. These modifications are described in the appendix of his 1981a paper. Source: Robert Shiller.

B. Data Set 2: Value-Weighted NYSE Index (Annual 1926–1981)

P_t represents the January value-weighted NYSE stock price divided by the January

producer price index (PPI). D_t represents total dividends for the year accruing to the portfolio represented by the stocks in the index, divided by the annual average PPI. In the Marsh and Merton (1987) analysis, P_t represents December rather than January prices, but January prices are considered here to remain consistent with the Shiller dating scheme. Source: Center for Research in Security Prices (CRSP; University of Chicago) data set.

C. Data Set 3: Modified Standard and Poor's Series (Annual 1871–1985)

The price series $\{P_t\}$ is Standard and Poor's monthly composite stock price index for January, divided by the producer price index (January PPI starting in 1900, annual average PPI before 1990 scaled to 1.00 in the base year 1979). The dividend series $\{D_t\}$ represents the total dividends for the calendar year accruing to the portfolio represented by the stocks in the index, divided by the average PPI for the year. These data differ slightly from Shiller's (1981a) numbers; some adjustments have been made to correct minor errors in the original data, and the series have been updated to 1985. Source: Robert Shiller.

REFERENCES

- Campbell, John Y. and Shiller, Robert J., (1987a) "Cointegration and Tests of Present Value Models," *Journal of Political Economy*, October 1987, 95, 1062–88.
- _____, and _____, (1987b) "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," unpublished manuscript, Princeton University, 1987.
- Cecchetti, Stephen G., Lam, Pok-Sang and Mark, Nelson C., "Mean Reversion in Equilibrium Asset Prices," *American Economic Review*, June 1990, 80, 398–418.
- Christiano, Lawrence J. and Eichenbaum, Martin, "Unit Roots in GNP: Do We Know and Do We Care?" NBER (Cambridge, MA) Working Paper No. 3130, 1989.
- Cochrane, John H., "How Big is the Random Walk in Real GNP?" *Journal of Political*

Economy, October 1988, 96, 893-920.

DeJong, David N., Nankervis, John C., Savin, N. E. and Whiteman, Charles H., (1989a) "Integration Versus Trend-Stationarity in Macroeconomic Time Series," Department of Economics Working Paper No. 89-99, The University of Iowa, December 1989.

_____, _____, _____ and _____, (1989b) "Unit Root Tests and Coin Tosses in Macroeconomic Time Series," Department of Economics Working Paper No. 89-14, The University of Iowa, June 1989.

_____ and Whiteman, Charles H., "Trends and Cycles as Unobserved Components in Real GNP: A Bayesian Perspective," *Proceedings of the American Statistical Association, Business and Economics Section*, 1989, pp. 63-70.

_____ and _____, "More Unsettling Evidence on the Perfect Markets Hypothesis: Trend-Stationarity Revisited," Department of Economics Working Paper No. 90-14, The University of Iowa, May 1990.

_____ and _____, "Trends and Random Walks in Macroeconomic Time Series: A Reconsideration Based on the Likelihood Principle," *Journal of Monetary Economics*, 1991, forthcoming.

Dickey, David A. and Fuller, Wayne A., "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root," *Econometrica*, July 1981, 49, 1057-72.

Fama, Eugene F. and French, Kenneth R., "Permanent and Temporary Components of Stock Prices," *Journal of Political Economy*, April 1988, 96, 246-73.

Flavin, Marjorie A., "Excess Volatility in the Financial Markets: A Reassessment of the Empirical Evidence," *Journal of Political Economy*, December 1983, 91, 929-56.

Fuller, Wayne A., *Introduction to Statistical Time Series*, New York: Wiley, 1976.

Geweke, John, "Exact Inference in the Inequality Constrained Normal Linear Regression Model," *Journal of Applied Econometrics*, April 1986, 1, 127-41.

_____, "The Secular and Cyclical Behavior of Real GDP in Nineteen OECD Countries, 1957-1983," *Journal of Business and Economic Statistics*, October 1988, 6, 479-88.

_____, (1989a) "Exact Predictive Densities for Linear Models with ARCH Disturbances," *Journal of Econometrics*, January 1989, 40, 63-86.

_____, (1989b) "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, November 1989, 57, 1317-39.

Kleidon, Allan W., "Variance Bounds Tests and Stock Price Valuation Methods," *Journal of Political Economy*, October 1986, 94, 953-1001.

Kloek, Teun and van Dijk, Herman K., "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica*, January 1978, 46, 1-19.

Leamer, Edward E., *Specification Searches*, New York: Wiley, 1978.

Lo, Andrew W. and MacKinlay, A. Craig, "Stock Market Prices Do Not Follow Random Walks: Evidence From a Simple Specification Test," *Review of Financial Studies*, Spring 1988, 1, 41-66.

Mankiw, N. Gregory, Romer, David and Shapiro, Matthew D., "An Unbiased Reexamination of Stock Market Volatility," *Journal of Finance*, July 1985, 40, 677-87.

Marsh, Terry A. and Merton, Robert C., "Dividend Variability and Variance Bounds Tests for the Rationality of Stock Market Prices," *American Economic Review*, June 1986, 76, 483-98.

_____ and _____, "Dividend Behavior for the Aggregate Stock Market," *Journal of Business*, January 1987, 60, 1-40.

Perron, Pierre, "Trends and Random Walks in Macroeconomic Time Series: Further Evidence from a New Approach," *Journal of Economic Dynamics and Control*, June/September 1988, 12, 297-332.

Phillips, Peter C. B., "Time Series Regressions with a Unit Root," *Econometrica*, March 1987, 55, 277-302.

_____, "Regression Theory for Near-Integrated Time Series," *Econometrica*, September 1988, 56, 1021-43.

_____ and Perron, Pierre, "Testing for a Unit Root in Times Series Regression," *Biometrika*, 1988, 75 (2), 335-46.

Poterba, James M. and Summers, Lawrence H., "Mean Reversion in Stock Prices: Evi-

- dence and Implications," *Journal of Financial Economics*, October 1988, 1, 27-60.
- Shiller, Robert J.**, "A Distributed Lag Estimator Derived from Smoothness Priors," *Econometrica*, July 1973, 41, 775-88.
- _____, (1981a) "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?" *American Economic Review*, June 1981, 71, 421-36.
- _____, (1981b) "The Use of Volatility Measures in Assessing Market Efficiency," *Journal of Finance*, May 1981, 36, 291-304.
- Sims, Christopher A.**, "Bayesian Skepticism on Unit Root Econometrics," *Journal of Economic Dynamics and Control*, June/September 1988, 12, 463-75.
- _____, and **Uhlig, Harald**, "Understanding Unit Rooters: A Helicopter Tour," Federal Reserve Bank of Minneapolis Institute for Empirical Macroeconomics Discussion Paper No. 4, 1988.
- Theil, Henri and Goldberger, A. H.**, "On Pure and Mixed Statistical Estimation in Economics," *International Economic Review*, January 1961, 2, 65-78.
- West, Kenneth**, "A Note on the Power of Least Squares Tests for a Unit Root," *Economics Letters*, 1987, 24 (3), 249-52.
- _____, "Dividend Innovations and Stock Price Volatility," *Econometrica*, January 1988, 56, 37-61.
- Zellner, Arnold**, *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley, 1971.

A Note on Optimal Fines When Wealth Varies Among Individuals

By A. MITCHELL POLINSKY AND STEVEN SHAVELL*

An important result in the economic theory of enforcement is that, under certain circumstances, it is optimal to impose the highest possible fine—equal to an individual's entire wealth—with a relatively low probability of detection. The reasoning supporting this conclusion, which is usually attributed to Gary S. Becker (1968), is well known: if the fine is not at its highest level, enforcement costs can be reduced without affecting deterrence. This can be done by raising the fine to its highest level and lowering the probability of detection proportionally, so that the expected fine—and thus deterrence—is unchanged. Hence, according to this argument, it cannot be optimal for the fine to be less than an individual's wealth.

It is puzzling, of course, that this result differs so much from reality. Fines equal to an individual's wealth hardly ever are imposed. Several explanations have been offered to reconcile Becker's theory with this fact. For example, it has been shown that, if individuals are risk-averse, fines less than their wealth generally are optimal; and it has been argued that it is inequitable to impose fines greatly exceeding the harm done.¹ This note provides a new explanation of why fines are limited.

*Stanford Law School, Stanford, CA 94305, and Harvard Law School, Cambridge, MA 02138, respectively. Both authors also are research associates of the National Bureau of Economic Research. Polinsky's research was supported by the John M. Olin Program in Law and Economics at Stanford Law School. Shavell's research was supported by the National Science Foundation (grant SES-8821400). Helpful comments were provided by Louis Kaplow and two anonymous referees.

¹In a previous article (Polinsky and Shavell, 1979), we develop the risk-aversion explanation, and R. A. Carr-Hill and N. H. Stern (1979 pp. 281–95) discuss the equity explanation (as well as several others). In addition, George J. Stigler (1970) proposes an explana-

We will demonstrate that, if the wealth of individuals varies, as is obviously realistic, the optimal fine is less than the wealth of the highest-wealth individuals and may be less than the wealth of most individuals. In other words, the optimal fine is such that only relatively low-wealth individuals pay everything they have; all other individuals pay the fine, which is less than their wealth.

To understand our conclusion, consider why the argument associated with Becker cannot be applied when wealth varies. Suppose that the fine is less than the wealth of the highest-wealth individuals. If the fine is raised and the probability of detection is lowered proportionally, it is true that those who can pay the higher fine are deterred to the same extent. However, those who cannot pay the higher fine are deterred less. As a consequence, it generally is not optimal to raise the fine to the highest possible level.

For example, a fine of \$100 for speeding may be optimal because many drivers may have so little in savings that it would be difficult to collect more from them. If a much larger fine were imposed with a much smaller probability, these drivers would be inadequately deterred. Thus, a fine of \$100 may be optimal, which would mean that all speeding drivers with wealth exceeding \$100 would pay a fine that is less than their wealth.

I. Analysis and Example

In the model, risk-neutral individuals contemplate whether to commit a harmful act. Each individual is identified by the ben-

tion based on marginal deterrence, and Shavell (1991) presents an explanation based on the assumption that the probability of detection is the same for different types of harm. See also Richard A. Posner (1986 pp. 205–12).

efit he would obtain from committing the act and by his level of wealth. If an individual commits the harmful act, he will be made to pay a fine with some probability; this probability is determined by the enforcement expenditures of the state.

The following notation will be used:

- h = harm caused if the harmful act is committed ($h > 0$);
- b = benefit from committing the harmful act ($b \geq 0$);
- $r(b)$ = probability density of b ($r > 0$ for all $b \geq 0$);
- w = wealth of an individual ($w \geq 0$);
- $s(w)$ = probability density of w ($s > 0$ for all $w \geq 0$);
- $f(w)$ = fine for committing the harmful act for an individual whose wealth is w ($0 \leq f(w) \leq w$);²
- c = enforcement costs of the state ($c \geq 0$);
- $p(c)$ = probability of detection ($p'(c) > 0$, $p''(c) < 0$).

The probability of detection is assumed to be the same for individuals of different wealth. This assumption is crucial, as will be commented upon below. Also, the distribution of benefits is assumed to be the same for different levels of wealth; this assumption is not essential.

Social welfare is the sum of the benefits obtained by individuals who commit the harmful act, less the harm done and less enforcement costs. To determine social welfare, observe that an individual will commit the harmful act if and only if³

$$(1) \quad b \geq pf(w).$$

Hence, social welfare is

$$(2) \quad \int_0^\infty \int_{pf(w)}^\infty (b - h)r(b) db s(w) dw - c.$$

²Implicit in the assumption that $f(w) \leq w$ is the further assumption that an individual's wealth does not include the benefit he obtains from committing the harmful act. The latter assumption is made only for convenience.

³The assumption that an individual commits the act when $b = pf(w)$ is immaterial.

Let us first determine the optimal fine, $f^*(w)$, assuming that the probability of detection, p , is positive. Clearly, given p and any w , $f^*(w)$ is the f that maximizes

$$(3) \quad \int_{pf}^\infty (b - h)r(b) db.$$

The derivative of (3) with respect to f is $p(h - pf)r(pf)$, which is positive for $f < h/p$, 0 at $f = h/p$, and negative for higher f . Thus, the optimal f equals h/p if h/p is feasible, that is, if $h/p \leq w$; otherwise, the optimal f is w . In other words, $f^*(w) = \min(h/p, w)$.

This result can be restated as follows. *The optimal fine equals an individual's wealth for every individual with wealth less than h/p ; for all other individuals, who have higher wealth, the optimal fine is h/p , which is less than their wealth.* Equivalently, the optimal fine is h/p for everyone, but those who cannot pay this amount pay what they can. Note that those who can pay h/p are optimally deterred (i.e., act in the first-best manner) since their expected fine equals the harm caused.⁴

Next consider the optimal probability of detection. Because $f^*(w) = \min(h/p, w)$, (2) can be rewritten as

$$(4) \quad \int_0^{h/p} \int_{pw}^\infty (b - h)r(b) db s(w) dw + \int_{h/p}^\infty \int_h^\infty (b - h)r(b) db s(w) dw - c.$$

The first term relates to individuals who pay their wealth w when fined because they cannot pay h/p ; the second term relates to individuals who have wealth of at least h/p

⁴In Polinsky and Shavell (1984 pp. 96–7), we briefly considered optimal fines when there are two types of individuals who differ in terms of wealth. It was shown there that the optimal fine for an individual in the low-wealth group is equal to his wealth and that the optimal fine for an individual in the high-wealth group is larger but not necessarily equal to his wealth. The analysis here generalizes that result and is consistent with it.

and who therefore pay h/p and are optimally deterred.

Setting the derivative of (4) with respect to c equal to zero gives the relevant first-order condition,

$$(5) \int_0^{h/p} p'(c)(h - pw)wr(pw)s(w)dw = 1.$$

The left-hand side is the marginal benefit of raising c : some individuals with wealth less than h/p are underdeterred since they cannot pay h/p ; by raising c , p is raised, and more such individuals are deterred; at the margin, there is a social gain of $h - pw$ from deterring an individual with wealth w . The right-hand side is the marginal cost of raising c , namely 1. The optimal p is determined implicitly by the optimal choice of c from (5).

The preceding results can be illustrated by a simple example. Suppose that the harm h if the harmful act is committed is \$20, that the benefit b from committing the act is uniformly distributed between \$0 and \$25, that the wealth w of individuals is uniformly distributed between \$0 and \$100,000, and that the probability of detection p as a function of enforcement costs c is given by $75c$. Then it can be shown that the optimal probability of detection p^* is 0.2 and that the optimal fine f^* is \$100 (as expected, $p^*f^* = h = \$20$).⁵ Thus, everyone with wealth less than \$100 pays his wealth, while everyone with wealth greater than this level pays the fine of \$100. Given the assumption that wealth is uniformly distributed between \$0 and \$100,000, the optimal fine is less than the wealth of more than 99 percent of the population and is less than 1 percent of the wealth of the highest-wealth individuals.

II. Comments

(a) As noted above, the assumption that the probability of detection is the same for individuals with different wealth is central

to our results. If the probability could be chosen independently for individuals with different levels of wealth, then, for each level of wealth w , the optimal fine would be the entire wealth w . The reason is that Becker's argument would apply for each w : if $f^*(w) < w$, then by raising f to w and lowering p from $p^*(w)$ to the p such that $pw = p^*(w)f^*(w)$, deterrence would not be affected, but enforcement costs would fall.

However, for many harmful acts, the probability of detection does seem to be largely independent of wealth. This may be because it is difficult to vary enforcement effort with respect to individuals' wealth. For example, there obviously are limits to our ability to make the probability of detecting traffic violations depend on the wealth of drivers. In these kinds of circumstances, a fine less than the wealth of many individuals generally will be optimal for the reasons explained in this note.

(b) The assumption that the benefits obtained by individuals who commit the harmful act are included in social welfare is necessary for our conclusions. If individuals' benefits did not count in social welfare, it would be desirable to deter all harmful acts. Consequently, it would always be optimal to impose the highest possible fine, that is, a fine equal to wealth.

(c) If the sanction were imprisonment rather than a fine, the results would be analogous to those discussed here. Suppose individuals vary in terms of their age, and therefore in terms of the maximum imprisonment term that can be imposed on them. Then the analogue to our result would be that the optimal imprisonment term generally would be less than the maximum imprisonment term that could be imposed on the youngest offenders.

REFERENCES

- Becker, Gary S., "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, March/April 1968, 76, 169-217.
- Carr-Hill, R. A. and N. H. Stern, *Crime; The Police and Criminal Statistics*, London: Academic Press, 1979.

⁵The optimal probability is determined by solving (5), which reduces in the example to $0.00000711/c^2 = 1$.

- Polinsky, A. Mitchell and Steven Shavell, "The Optimal Tradeoff between the Probability and Magnitude of Fines," *American Economic Review*, December 1979, 69, 880-91.
- ____ and _____, "The Optimal Use of Fines and Imprisonment," *Journal of Public Economics*, June 1984, 24, 89-99.
- Posner, Richard A., *Economic Analysis of Law*, 3rd Ed., Boston: Little, Brown, 1986.
- Shavell, Steven, "Specific versus General Enforcement of Laws," *Journal of Political Economy*, 1991, forthcoming.
- Stigler, George J., "The Optimum Enforcement of Laws," *Journal of Political Economy*, March/April 1970, 78, 526-36.

Standards versus Standards: The Effects of Different Pollution Restrictions

By GLORIA E. HELFAND*

When economists refer to pollution standards, they almost universally mean uniform restrictions on pollution emissions (e.g., William J. Baumol and Wallace E. Oates, 1975 Ch. 13; Susan Rose-Ackerman, 1973; David Besanko, 1978). However, in practice, standards take many forms: not only emissions restrictions, but restrictions on pollution per unit of output or per unit of an input, restrictions on the use of a polluting input, or mandated use of a particular pollution-control technology. These specifications of regulations will have a variety of effects on a firm's resource-allocation decisions. For instance, allowing pollution per unit of output as the standard instead of directly restricting the level of pollution gives the firm some choice: in addition to reducing pollution, it can increase its output to "dilute" the pollution.

This paper examines the effects of five different forms of pollution standards on input decisions, the level of production, and firm profits. The results show that the different standards, by providing firms with different incentives, change the firm's allocation decisions and affect the relative profitability of the standards. Section I reviews the existing literature on the incentives of different instruments. The following section describes the model to be used. Five different forms of pollution constraints are individually examined in Section III using a graphical approach. Their relative effects,

as well as some special cases, are assessed in Section IV. Finally, Section V raises the case of different standards when firms are not identical.

I. Different Standards and What Is Known of Them

The forms of standards used in actual regulations vary tremendously. A review of various state and federal regulations would show that standards are set in terms of such units as the total level of emissions per unit time, the amount of emissions per unit of output or input, requirements for certain pollution-control technologies, and restrictions on polluting inputs (Clifford S. Russell et al., 1986 pp. 17-19; Helfand, 1988 Ch. 2).

While the pollution-control literature includes analysis involving many of these forms, the correspondence between frequency of use in the literature and frequency of use in the standards is not perfect. For instance, while the literature includes examples of standards expressed as restrictions on output (e.g., James M. Buchanan and Gordon Tullock, 1975), no example of such a standard was found in any regulations reviewed. Many national regulations are set in terms of emissions per unit of output or input, but relatively few articles use that form of standard (Eithan Hochman and David Zilberman [1978] is an exception). Standards with such varied forms must inevitably provide firms with different incentives. Studies that examine the effects of a particular standard may not generalize to other forms of standards.

A few studies have examined pieces of this problem, but no one study has looked in a general way at the range of standards in use. Jon D. Harford and Gordon Karp (1983) compared the efficiency of standards expressed as pollution per unit of output

*Assistant Professor, Department of Agricultural Economics, University of California, Davis, Davis, CA 95616. I thank Peter Berck, Larry Karp, Jonathan Rubin, David Zilberman, Susanne Scotchner, seminar participants at the University of Illinois, the University of California at Berkeley, the University of California at Davis, and Duke University, and two anonymous referees for their helpful comments. This is Giannini Foundation Paper No. 950.

and pollution per unit of input. Their model, which held output and total emissions fixed, found that the pollution-per-output standard was more efficient because it least distorted the input mix.

Vinod Thomas (1980) compared the welfare costs of an emissions standard to several other forms of regulation for the steel industry near Chicago: a set level of pollution-control expenditures, restricting the polluting input (fuel), and restricting emissions as a function of fuel. The welfare costs of these other policies, except the input restriction, were estimated to be 30–40 percent higher than those of the emissions standard, though only 1–1.4 percent of the value of steel output; direct restriction of the fuel input had a welfare cost of about 30 percent of the value of output. Thomas concluded that the choice of a standard could have a significant effect on the efficiency costs of environmental regulation.

Besanko (1987) analyzed the effects of restricting pollution (a “performance” standard) versus mandating a specific level of pollution-control technology (a “design” standard). For N symmetric Cournot firms, the performance standard resulted in higher profits for the individual firms, but output was higher under the design standard. If social welfare is defined as the sum of producer and consumer surplus, neither standard unambiguously produced higher welfare than the other: the greater output under the design standard enhanced consumer surplus, even as it reduced producer surplus.

The articles by Harford and Karp and by Besanko look only at subsets of possible standards, while the article by Thomas examines the effect of a range of standards only for one industry. Because the models used in these studies have different assumptions, the results are not directly related. The present study reviews a range of standards in the context of one model to compare their effects unambiguously.

II. The One-Firm Model

The model used here involves one firm, facing a horizontal output demand curve

and using two inputs, x_1 and x_2 , with horizontal supply curves. While the assumption that there are only two inputs is a simplification, made to keep the comparative statics and the graphical presentation simple, the problem is likely to generalize to N inputs with few differences. The assumption of a horizontal output demand curve is more limiting. Made to keep the problem tractable and to permit a graphical presentation, it is realistic only for a good whose world price is unaffected by production in this country. It deserves to be lifted in future work.

The firm is assumed to produce an output f using the production function $f(x_1, x_2)$. It is initially assumed that $f_i > 0$ for $i = 1, 2$,¹ that is, both inputs contribute to production. Additionally, $f_{ii} < 0$ for both i , and the Hessian of the production function is negative definite; that is, both inputs and the production function are subject to diminishing marginal returns.

The firm also produces pollution $A(x_1, x_2)$. It is initially assumed that input 1 increases pollution (i.e., $A_1 > 0$), while input 2 decreases pollution ($A_2 < 0$). For instance, input 1 can be thought of as coal used in electricity generation, and input 2 as water used both to turn the turbines for electricity and to cool thermal discharges.

The firm is assumed to maximize profits while facing an output price p and input prices w_1 and w_2 , corresponding to inputs 1 and 2, respectively. They are assumed to be fixed for this analysis.

Before a pollution-control constraint is imposed, the firm is assumed to maximize profits, π . The unconstrained problem becomes

$$(1) \quad \max_{x_1, x_2} \pi = pf(x_1, x_2) - w_1x_1 - w_2x_2$$

which results in a solution $x^0 = (x_1^0, x_2^0)$; it is unique, since the production function is strictly concave.

¹Subscripts in this and future cases denote derivatives with respect to i .

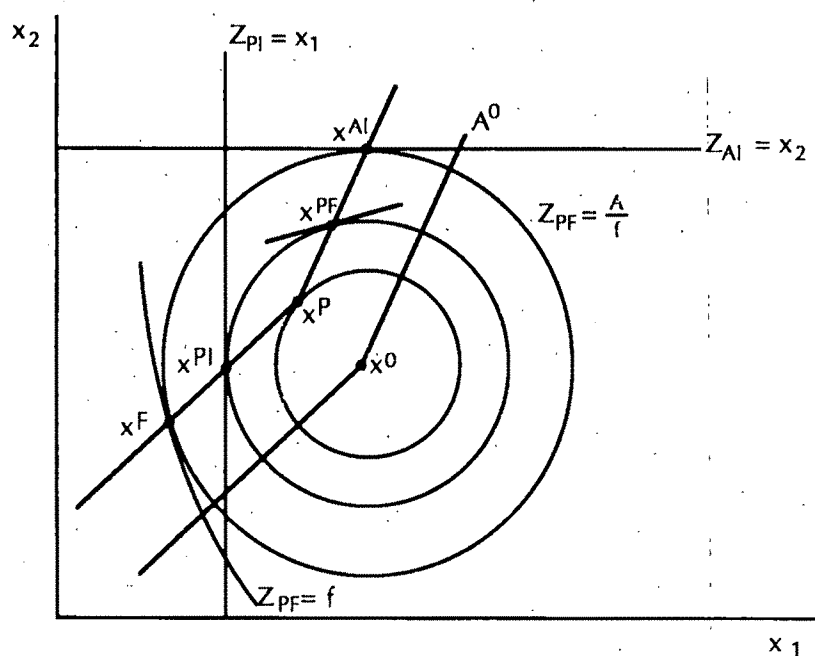


FIGURE 1. EFFECTS OF DIFFERENT POLLUTION-CONTROL STANDARDS

This result can be displayed on a diagram like one used for isoquant analysis, such as Figure 1. The polluting input, x_1 , is measured along the horizontal axis, while the pollution-abating input, x_2 , is measured along the vertical axis. The profit-maximizing combination of these inputs is found at x^0 , which implicitly lies at the tangency of an isoquant $f^0 = f(x_1^0, x_2^0)$ with an isocost line whose slope is $-w_1/w_2$.

Because of the uniqueness of this maximum, isoprofit contours (i.e., contours representing input combinations yielding the same level of profits) can be drawn on the diagram as well. That these contours are convex rings can be derived from the assumed strict concavity of the production function. The maximum-profits point x^0 can be considered the top of a profits hill. Iso-profit contours away from this point are decreasing levels of profits with movement down the profits hill.

An isopollution line (i.e., a line representing input combinations giving a fixed level

of pollution) can be added to this diagram. Totally differentiating the function $A^0 = A(x_1, x_2)$ gives the slope of this line as $dx_2/dx_1 = -A_1/A_2$, which is positive since $A_1 > 0$ and $A_2 < 0$. In other words, as use of the polluting input increases, use of the abating input must also increase in order to maintain a constant level of pollution. Isopollution contours above and to the left of another contour represent lower levels of pollution: for a given amount of the pollution-abating input, pollution decreases as use of the polluting input decreases. In Figure 1, the initial level of pollution is represented by the line $A^0 = A(x_1^0, x_2^0)$. The line A^1 , above and to the left of the original contour, represents a lower level of pollution.

III. The Constraints

Five different kinds of pollution-control standards will be examined here: a fixed level of emissions, a fixed level of emissions

per unit of output, a fixed level of emissions per unit of an input, a fixed level of output, and a fixed level of an input.² In two of these cases (pollution per input and fixing the level of an input), two subcases will be analyzed, reflecting the different inputs used here. These formulations include most of the standards analyzed by Thomas (1980), Harford and Karp (1983), and Besanko (1987), as well as others used in either the pollution-control literature or environmental regulation.

One way to analyze the effects of each constraint relative to the baseline of no regulation is to set up the profit-maximization problem, differentiate totally the first-order conditions, and use comparative statics to analyze the effects of the constraint on the level of input use, output, profits, and pollution. Identical results can be derived from a graphical analysis. Since a graphical approach is easier to interpret, it will be used to present the results; the comparative-statics analysis can be obtained from the author.

A. Standard as a Set Level of Emissions

Let Z_P be the numerical standard set when emissions are regulated by the amount of total pollution permissible in a certain period of time. It can be represented as a constraint on the profit function with the form $A \leq Z_P$.

Figure 1 presents the graphical interpretation of this problem. The point $x^0 = (x_1^0, x_2^0)$ is the previously described unconstrained-maximum-profits locus. The addition of the constraint requires that the firm

choose inputs that will place it on a leftward isopollution line, such as A^P . Since the firm will nevertheless seek to maximize profits, it will choose the point of the new isopollution line that achieves the highest level of profits: the point where the isopollution line is tangent to the highest attainable isoprofit line. This point is represented by $x^P = (x_1^P, x_2^P)$.

In general, as Figure 1 shows, use of the polluting input should decrease, and use of the pollution-abating input should increase. These results are expected, since the purpose of the constraint is to reduce pollution. However, use of the polluting input may actually increase, or use of the abating input may decrease. These seemingly perverse results depend on the actual shape of the isoprofit contours: the tangency between the isoprofit and isopollution line may lie above and to the right, or below and to the left, of x^0 . Mathematically, these results depend on the sign and magnitude of the f_{12} term, the effect on the marginal product of input 1 as use of input 2 changes. All that is known theoretically about this term is that $f_{11}f_{22} > (f_{12})^2$ for a strictly concave production function. This information is not sufficient to determine more about the effects of this constraint on input use.

Because use of one input is increasing and use of the other is decreasing, the effects on production are ambiguous. Profits are unambiguously lower under this standard, since the firm can always do better in an unconstrained situation than in a constrained one. Pollution, by definition of the constraint, is reduced.

B. Standard as Emissions per Unit of Output

Let Z_{PF} be the standard expressed as a set level of pollution per unit of output, or $A/f \leq Z_{PF}$. Totally differentiating the constraint equation gives the slope of the constraint line, $dx_2/dx_1 = -(fA_1 - f_1A)/(fA_2 - f_2A)$. The denominator is negative according to the assumptions on signs; $fA_1 - f_1A$ is positive if $A_1/(A/x_1) > f_1/(f/x_1)$. If A is convex in input 1 (indicating that

²These standards are not always as different as they first appear. For instance, a regulatory agency may develop a standard based on a particular cleanup technology (fixing an abating input) but express that standard in other units, such as pollution per unit of output. Indeed, the technology standard may de facto be mandated by the pollution-per-output standard. This analysis assumes that expression of the standard as pollution per unit gives the firm the choice of using alternative approaches.

pollution increases more rapidly as use of the polluting input increases), the marginal amount of pollution is greater than the average amount at any given point, and $A_1/(A/x_1)$ is greater than one. On the other hand, f is assumed to be concave in input 1; by analogous reasoning, the marginal product is less than the average product, and the ratio is less than 1. Therefore, $fA_1 - f_1A$ is positive, and the slope of this standard is positive.

The slope of this constraint line can be compared to the slope of the isopollution contour by subtraction:

$$\frac{dx_2}{dx_1}(Z_{PF}) - \frac{dx_2}{dx_1}(Z_P) = \frac{(f_1A_2 - f_2A_1)}{(fA_2 - f_2A)A_2} < 0$$

by the assumed signs of these terms. Though this line slopes upward, it slopes upward at a lesser angle than does the isopollution contour.

This new constraint is also illustrated in Figure 1. If the standard is normalized to achieve the same level of pollution as standard P, its constraint line must intersect the isopollution line at the constraint's tangency with the highest possible isoprofit contour. This fact, plus the fact that its slope is shallower than that of the isopollution line, can establish the location of this constraint. If it were along the same isoprofit contour as is standard P, then the tangency would be to the right of the tangency for standard P, x^P ; however, if it were along that isoprofit contour, then it could not be on the same isopollution line, since it would lie above that isoprofit contour except at x^P . Therefore, the combination of inputs that maximizes profits while achieving standard PF, $x^{PF} = (x_1^{PF}, x_2^{PF})$, is above and to the right of the combination for standard P along the isopollution line A^P .

In the "normal" case, use of the polluting input should drop, and use of the abating input should increase. As with the pollution standard, though, it is possible that use of

the polluting input may increase or that use of the abating input may decrease. These results depend, as with the pollution constraint, on the sign and the magnitude of the f_{12} term. The effects on production remain ambiguous; profits unambiguously decrease. Finally, the effects of this standard on pollution levels are somewhat ambiguous, depending on the sign and magnitude of the f_{12} term: if production increases more rapidly than pollution, then this constraint could lead to the perverse result that pollution increases with its imposition.

C. Standard as Emissions per Unit of a Specified Input

The mathematical representation of this standard is $A/x_j \leq Z_{PJ}$, $J = 1$ or 2 , where J is the subscript for the input in terms of which pollution is measured. Two cases are possible here: regulating pollution per unit of a pollution-causing input, such as restricting the amount of sulfur dioxide emissions per ton of coal used for electricity; or regulating pollution per unit of a pollution-reducing input, such as regulating biological oxygen demand in water pollution per unit of water in a production process. Wesley A. Magat et al. (1986 p. 35) note that the Environmental Protection Agency (EPA) preferred a pollution-per-output standard to the latter form, out of concern that plants would meet the standard solely by dilution and not by cleanup.

Let PPI represent the situation in which pollution is regulated per unit of the polluting input, and let PAI represent pollution regulated in terms of the pollution-abating input. The slope for $Z_{PPI} = A/x_1$ is $-(A_1x_1 - A)/A_2x_1$ which is positive if $A_1 > A/x_1$. By the same reasoning used for $f_1A - fA_1$ in the pollution-per-output case above, the marginal pollution at any given level of inputs is greater than the average level if A is convex in the polluting inputs. Thus, $A_1x_1 - A$ is positive, and this slope is positive. For $Z_{PAI} = A/x_2$, the slope is $-A_1x_2/(A_2x_2 - A) > 0$. As with pollution per output, the slopes of these constraint lines can be compared with that of the

isopollution line:

$$(2) \quad \frac{dx_2}{dx_1}(Z_{PPI}) - \frac{dx_2}{dx_1}(Z_P) = \frac{A}{A_2 x_1} < 0$$

$$\frac{dx_2}{dx_1}(Z_{PAI}) - \frac{dx_2}{dx_1}(Z_P)$$

$$= \frac{AA_1}{A_2(A - A_2 x_2)} < 0.$$

Clearly the slopes of the constraint lines are less than the latter slope. However, no clear comparison of slopes is possible between these two standards or between either of these standards and the pollution-per-output standard: none of the following comparisons can be signed:

$$(3) \quad \frac{dx_2}{dx_1}(Z_{PPI}) - \frac{dx_2}{dx_1}(Z_{PAI})$$

$$= \frac{A(A_1 x_1 + A_2 x_2 - A)}{A_2 x_1(A_2 x_2 - A)}$$

$$\frac{dx_2}{dx_1}(Z_{PPI}) - \frac{dx_2}{dx_1}(Z_{PF})$$

$$= \frac{A[A_2(f - f_1 x_1) + f_2(A_1 x_1 - A)]}{A_2 x_1(f A_2 - f_2 A)}$$

$$\frac{dx_2}{dx_1}(Z_{PAI}) - \frac{dx_2}{dx_1}(Z_{PF})$$

$$= \frac{A[A_1(f_2 x_2 - f) + f_1(A - A_2 x_2)]}{(A_2 x_2 - A)(f A_2 - f_2 A)}.$$

The numerator of the slope difference between standards PPI and PAI would be zero if the pollution function had constant returns to scale (since, by Euler's law, $A = A_1 x_1 + A_2 x_2$ if the function has constant returns to scale); however, without assumptions about the returns to scale of this function, it cannot be signed. It nevertheless indicates that these returns determine any

differences between these functions: increasing returns makes the PAI standard line steeper, and decreasing returns make PPI have a steeper slope. Somewhat similarly, if both the production and the pollution functions have constant returns to scale, then the numerators of the comparisons between either pollution-per-input standard and the pollution-per-output standard would be zero as well.³ Thus, comparison of either standard with the constraint line for standard PF requires further knowledge of the relative returns to scale for both the pollution and the production functions.

Because the constraint lines of these three standards (pollution per unit of output and pollution per unit of either input) cannot be readily distinguished from each other, the remainder of this analysis will not distinguish among them. These three formulations of the standard will be referred to as the "dilution" standards, since they all involve measuring pollution diluted by either output or input. The point x^P in Figure 1 thus reflects the effects of standards PPI and PAI as well as it reflects the effects of standard PF.

D. Standard as a Set Level of Total Output

This standard forces the firm to reduce output; input substitution alone is inadequate. The constraint is now $f \leq Z_F$. This formulation of the standard mandates that the firm operate on an isoquant closer to the origin than the profit-maximizing isoquant. The firm will operate with the input mix given by the point of tangency of that downward-sloping isoquant with the highest possible isoprofit curve, represented by the point $x^F = (x_1^F, x_2^F)$. Because of the negative slope of the isoquant, this tangency must be below and to the left of x^P .

³With constant returns, $f - f_1 x_1 = f_2 x_2$, and $A_1 x_1 - A = -A_2 x_2$; the numerator of $(dx_2/dx_1)(Z_{PPI}) - (dx_2/dx_1)(Z_{PF})$ becomes zero. A similar analysis can be done for the terms in the comparison of standard PAI and standard PF.

Use of both inputs will generally decrease, though as before, exceptions are possible, depending on f_{12} . By definition, output decreases, and profits decline as usual, because the firm is constrained away from the optimum.

It should be noted that pollution need not decrease with this standard: the new optimal input mix could end up on an isopollution curve either to the left or to the right of the initial curve, depending on whether the firm uses the polluting input or the abating input more intensively as production is reduced. It is assumed in the drawing of Figure 1 that the lower level of pollution will be achieved by a reduction in output.

E. *Standard as a Set Amount of a Specified Input*

This standard takes two forms. A maximum can be set on the use of a polluting input (standard PI); alternatively, imposing a minimum level on the use of a pollution-abating input (standard AI) captures the effect of imposing a particular pollution-control technology on a firm. Standard PI will be represented as $x_1 \leq Z_{PI}$, and standard AI will be represented as $x_2 \geq Z_{AI}$.

Graphically, the limitation on use of the polluting input is a vertical line at the chosen level of input 1; the requirement for a minimum use of the pollution-abating input results in a horizontal line at the chosen level of input 2. The vertical slope of standard PI places its optimum, x^{PI} , along A^P between x^P and x^F . The zero slope of standard AI places x^{AI} above and to the right of x^{PF} along A^P . These are shown in Figure 1.

For case PI, the limit on use of the polluting input obviously causes a reduction in the use of input 1. The effect on input 2 is ambiguous, depending entirely on the sign of f_{12} . Production will be lower if use of input 2 is not increased much, is unaffected, or decreases. Profits are obviously reduced, as in all cases. Finally, as with the output standard, pollution could either increase or decrease with the formulation of the standard, depending on the slope of the isopol-

lution line and the shape of the isoprofit contours.

For case AI, use of input 1 may either increase or decrease, depending on the sign of f_{12} . If f_{12} is nonnegative or not very negative, then production will increase through use of this standard, in contrast to the effects of standard PI. Profits decrease as usual. Finally, pollution could either increase or decrease with the use of this standard.

In sum, these different forms of pollution constraints have some similar patterns, but each has its own peculiarities. While standard AI, the minimum requirement on the abating input, may actually increase output, standard PI, the maximum requirement on the polluting input, and standard F, the restriction on output, should decrease production. The other standards have ambiguous effects on output, since they all tend to decrease use of the polluting input and to increase use of the pollution-abating input increase. Pollution unambiguously decreases only for standard P and for the dilution and input standards if f_{12} is "small" (i.e., small enough that terms involving it do not affect the sign of the function as a whole) or zero. For standard F, the effects of the restrictions on pollution are completely ambiguous. For all standards except AI, use of the polluting input is expected to drop as the pollution-control constraint is tightened; for all standards except F and PI, use of the pollution-abating input should increase.

IV. Comparisons of the Different Standards

The previous section describes the results of the different standards as they are imposed. As noted, most of the results have the same sign and are therefore initially indistinguishable. However, if the standards are normalized to achieve the same level of pollution, as they have been in Figure 1, comparisons among the standards can easily be made.

This normalization is necessary because the standards differ in form—indeed, even in units of measurement. A change of a given magnitude in one standard is not

equivalent to a change of the same magnitude in another standard. For instance, changing Z_P by one unit will change the level of pollution by one unit; however, changing Z_{PF} by one unit will result in a one-unit change in pollution per unit of output, and the results are not obviously comparable. The standards have to be normalized to reflect a common level of change. The normalization used here is to make each standard cause an identical change in the amount of pollution. The standards have been ordered along the isopollution line A^P by comparing the slopes implied by the different standards, as already discussed.

This procedure reveals a strict ordering among the standards for levels of input use and levels of output as well as some information on relative profits. The standard that most reduces input use and output levels is standard F, the restriction on output, followed by standard PI, the restriction on the polluting input. Standard PI gives the firm higher profits than does standard F. Although standard P, the restriction on pollution itself, gives the highest level of profits among any of these standards, it has lower levels of input use and output than do the dilution standards, which in turn have lower levels of input use and output (though higher profits) than result from mandating a minimum amount of the pollution-abating input.⁴

In sum, even if the individual comparative-statics results for almost all the standards are ambiguous, the orderings among the standards are clear. Though a direct restriction on pollution offers the highest profit to a firm, mandating a minimum amount of the pollution-abating input leads to the highest production, while restricting output or the polluting input causes lower levels of production.

⁴The relative levels of profits between standards F and PF, standards F and AI standards PI and PF, and standards PI and AI, cannot be determined from this analysis. On the diagram, standards F and AI and standards PI and PF are drawn on the same isoprofit contours only to keep the diagram simple.

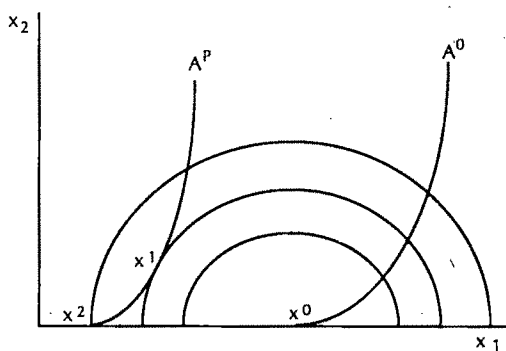


FIGURE 2. EFFECTS OF DIFFERENT STANDARDS WHEN THE ABATING INPUT DOES NOT CONTRIBUTE TO PRODUCTION

Special Cases

These findings depend on some specific assumptions underlying the model. In particular, the assumption that the second input contributes both to production and to pollution abatement affects the differences among the standards studied here. Analysis of two special cases of the model provides some interesting contrasts with these findings.

(i) $f_2 = 0$.—In this case, it is assumed that the pollution-abating input does not contribute to production; rather, it only affects pollution. For instance, x_2 may be a pollution device attached to a smokestack that does not otherwise intervene in the production process. The isoquants are now vertical lines (since production only depends on input 1), with the unconstrained profit-maximizing point on the x_1 axis (i.e., where no use of x_2 is required, since x_2 does not contribute to production and is costly). As seen in Figure 2, assuming $f_2 = 0$ reduces the standards to two forms: x^1 , which represents the effects of a restriction of pollution (Z_P), of pollution per output (Z_{PF}), of pollution per input (Z_{PPI} or Z_{PAI}), or a required amount of the abating input (Z_{AI}); and x^2 , which results from either a restriction on the polluting input (Z_{PI}) or a restriction on output (Z_F) (since output results only from this input). Similarly to the

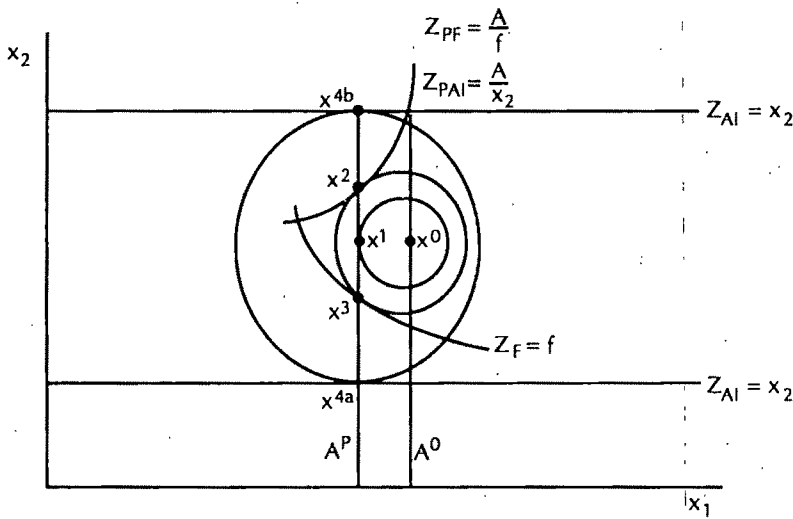


FIGURE 3. EFFECTS OF DIFFERENT STANDARDS WHEN THE SECOND INPUT DOES NOT ABATE POLLUTION

basic model, the standards yielding x^1 lead to higher levels of input use, output, and profits than the standards yielding x^2 .

(ii) $A_2 = 0$.—It is now assumed that the second input in the process does not abate pollution but has a neutral effect; the only way to reduce pollution is to reduce the level of the polluting input. As shown in Figure 3, the isoprofit contours are the same as in the basic model; however, the isopollution lines are vertical, since they are affected only by use of x_1 . Now, with all standards normalized to achieve A^P , restricting pollution per unit of output (Z_{PF}) and restricting pollution per unit of the second input (Z_{PAI}) have the same effect, a high level of production (x^2); restricting pollution (Z_P), restricting pollution per unit of the polluting input (Z_{PPI}), and restricting the polluting input (Z_{PI}) lead to the highest profits (x^1); and restricting output (Z_F) leads to lower levels of production (x^3). In interesting contrast to the basic case, a positive f_{12} (at point x^{4a}) causes a mandate on the second input (Z_{AI}) to have the most severe output reduction: since x_2 now does not affect pollution directly, it can only achieve the standard by inducing a reduction in use of x_1 . If f_{12} is negative (at x^{4b}),

this standard leads to the greatest increase in production, again by inducing a change in the level of x_1 .

These special cases demonstrate the dependence of these results on the assumptions of the model. If the abating input does not affect production (as would, for instance, a device attached only to a smokestack or an outfall pipe), then the standards collapse into two different levels of effects. If the only way of abating pollution is to reduce the level of the polluting input, then a range of standards is maintained, though it is not the same range as in the basic model. Any application of this model should consider the specific assumptions of the case to be analyzed, in order to account for these differences.

V. The Case of Heterogeneous Firms

The original analysis assumed that an individual firm was being regulated and that the different standards would all result in the same total level of pollution. While this analysis provides insights into the incentives of these different standards, it is not realistic. In fact, the EPA and state regulatory bodies usually set standards that apply uniformly to firms of a specific industry. Within

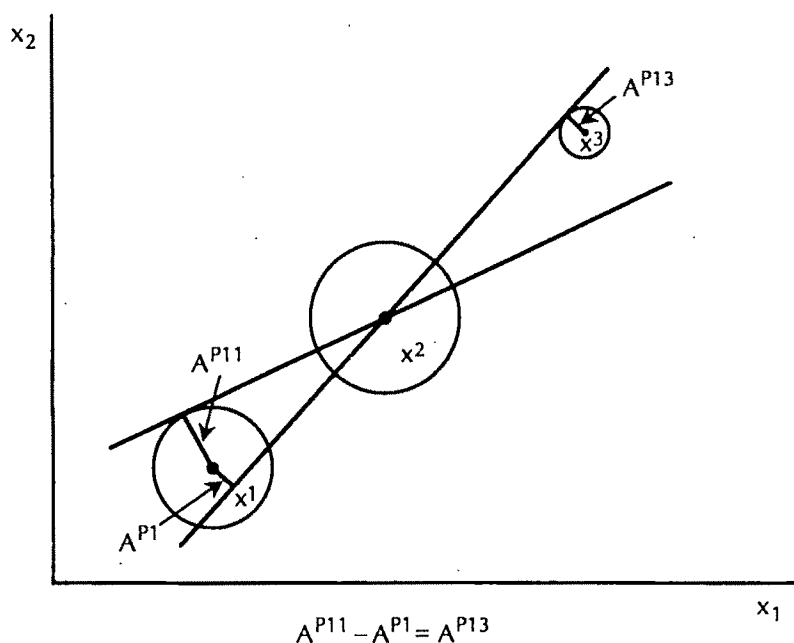


FIGURE 4. UNIFORM STANDARDS WITH HETEROGENEOUS FIRMS

an industry, individual firms are likely to vary by technology, size, or other factors. Given that only one form of standard is used to regulate an industry, it is unlikely that the results for the individual-firm case will carry over to the case of heterogeneous firms. This analysis will sketch some of the issues involved in choosing a standard for a number of different firms.

Assume that an industry contains three firms, all of which have the characteristics of the individual firm in the previous analysis, but which differ in optimal scale. In other words, firm 1's profit-maximizing level of output (x^1) uses less x_1 and less x_2 , and produces less output, than firm 2's profit-maximizing level of output (x^2), which in turn uses less x_1 and x_2 and produces less output than firm 3's profit-maximizing level of output, (x^3) (see Fig. 4).

Further assume that the pollution function $A(x_1, x_2)$ previously described is the same for all three firms. That is, only the inputs cause or abate pollution; the size of the firm itself does not have a direct effect on how much a firm pollutes. Thus, if firm 1

chose to operate with the same input mix as firm 2, it would pollute the same amount as firm 2. Under normal operating conditions, though, the level of pollution will change with the firm, since the firms use different levels of x_1 and x_2 .

This assumption simplifies a graphical analysis. Because most of the standards addressed in this study (in particular, the pollution standard, the pollution-per-input standards, and the input standards) rely only on the levels of the inputs and the pollution function; they do not depend on the level of output and, therefore, do not depend on the technology of the firm. As a result, one constraint can be drawn onto an isoquant diagram and will apply equally to all firms.⁵

⁵The pollution-per-output standard and the output standard will clearly have effects that depend on what technology a firm uses. For instance, these firms may have isoquants that cross each other. With these two standards, a single constraint line cannot be used, complicating a graphical analysis. For present purposes, these constraints will not be considered. While ignoring the output constraint does not have significant

Because the pollution-per-input standard⁶ is commonly used, while the pollution standard is heavily used in the economics literature, this analysis will focus on these two standards.

Firms are in compliance with the standards if they lie above or to the left of the constraint lines appropriate for the form of standard chosen. They initially violate the standards and must readjust their operations if they lie below or to the right of the constraint lines.

Figure 4 shows the constraint lines for these two standards. As in the single-firm analysis, the pollution constraint slopes upward, while the pollution-per-input standard slopes upward but more shallowly. (These relationships still hold, since both of these constraints depend only on the levels of input use.)

Where a firm lies on the diagram determines whether it has to adjust its production to comply with the constraints. In this diagram, firm 2 is in exact compliance with both standards; firm 1 satisfies the pollution standard but not the pollution-per-input standard; and firm 3 meets the pollution-per-input standard but not the pollution standard.

In the one-firm analysis, all the standards were oriented to produce the same total emissions from the one firm. Here, instead of making emissions per firm constant, the total level of pollution from all the firms will be held constant. This analysis will describe how to make such a normalization.

Let Z^P be the level of pollution specified by the pollution standard, and let A^{ij} represent the absolute value of the deviation from Z^P for firm j ($j=1,2,3$) under standard i ($i=P,PI$; where PI is pollution per unit of input for this analysis). Under the

pollution standard, firm 3 will have to reduce its pollution to Z^P to meet the standard, while firm 2 already pollutes exactly Z^P . Firm 1 already has lower emissions than this standard: it underpollutes by A^{P1} . Thus, total pollution under the pollution standard (TP_P) is $TP_P = 3Z^P - A^{P1}$.

Under the pollution-per-input standard, firm 1 has to clean up to meet the standard, even with its relatively low emissions. In contrast, firm 3 satisfies the standard and does not have to abate its emissions, though it pollutes more than Z^P . Total emissions under this standard (TP_{PI}) are

$$(4) \quad TP_{PI} = 3Z^P - A^{PI1} + A^{PI3}.$$

For total emissions to be the same under both standards, then $TP_P = TP_{PI}$, or

$$(5) \quad A^{PI1} - A^{P1} = A^{PI3}.$$

This normalization is approximated in Figure 4. In general, an exact normalization would require more complete knowledge of how changing the level of pollution changes the distance between the isopollution lines.

The effects of the different standards on the levels of input use, output, and profits for an individual firm relative to the non-pollution baseline follow the patterns of the one-firm analysis, with a notable exception: the small firm is completely unaffected by the pollution standard, while the large firm must adjust its behavior; in contrast, the pollution-per-input standard forces the small firm to adjust its behavior and leaves the large firm unaffected.

Comparisons between the standards are more interesting, but they are also more difficult to make, since they depend heavily on the shapes of the isoprofit contours. In the one-firm analysis, the pollution-per-input standard uses more of both inputs than the pollution standard, resulting in more output, though it gives the firm lower profits. With heterogeneous firms, however, similar conclusions cannot be drawn. Indeed, virtually any change in aggregate input use is possible. For instance, the shallower slope of the pollution-per-input standard may appear to induce less reduction in the polluting input and more increase in the

policy results, since this standard is not commonly used, further analysis of the common pollution-per-output standard is desirable.

⁶As seen in the one-firm analysis, the effects of measuring pollution per unit of the abating input are difficult to distinguish from the effects of measuring pollution per unit of the polluting input. For the purposes of this analysis, no distinction is drawn between the two.

abating input for the small firm than the pollution standard induces for the large firm. However, the "perverse" cases of the one-firm analysis can reverse this appearance. If the large firm increases its use of the polluting input or the small firm reduces its use of the abating input, then the relative change in total use of the inputs under the two standards cannot be determined. The shapes of the isoprofit contours for the firms, as well as the exact slopes of the two constraints, will determine the relative effects.

Because input comparisons cannot be made definitively, the effects of the different standards on output cannot be assessed. Finally, the relative effects on aggregate profits cannot be determined. If one firm has a shallower "profit hill" than the other (i.e., if profits for one firm decrease less rapidly than for the other when the profit-maximizing input mix is no longer available), then that firm is less likely to be severely hurt by either pollution standard. There is no obvious reason for smaller firms to have steeper profit hills than larger firms; as a result, no clear comparison can be made.

Perhaps the major conclusion from Figure 4 is that the effects of the different standards depend on the size of the firm in question. If pollution increases with output, even if not in direct proportion, then the large firm is more likely to have to adjust to a pollution standard than will the small firm.⁷ In contrast, a pollution-per-input standard "dilutes" the effect of pollution increasing with scale, because use of the input also increases with scale; the large firm is thus less likely to be penalized under this standard than under the pollution standard. The small firm may violate the pollution-per-input standard, even though it pollutes relatively small amounts.

⁷As drawn, pollution does increase with output, since the expansion path for firm scale has a flatter slope than the isopollution line. For output to increase while pollution declines, the slope of this expansion path would have to be steeper than the slope of the isopollution line. While such a scenario is possible if the expansion path is relatively intensive in input 2, it is generally more likely that larger firms pollute more than smaller firms.

These two standards provide different incentives for firms entering and exiting the industry. If pollution does increase with the size of the firm, then the pollution standard will put more pressure on large firms; they are more likely to exit, and small firms are more likely to enter. In contrast, the pollution-per-input standard will lead to larger firms entering the industry and smaller firms exiting. Ironically, this standard could lead to increased pollution, since large firms can comply with the pollution-per-input standard and still pollute more than small firms. The PI standard would have to become increasingly strict to maintain a lower level of emissions, which would in turn induce new firms to be even larger. The uniform pollution standard would have to be adjusted as the number of firms in the industry changed, but it would obviously not lead to more pollution per firm.

VI. Summary and Conclusions

This study has reviewed the effects of five different forms of pollution-control standards. It has shown in the basic model that a direct restriction on pollution leads to the highest level of profits and efficiency when all pollution standards are set to achieve the same level of emissions; a mandate for use of a pollution-abating input leads to highest output, followed by pollution per unit of output or input; and a restriction on production most reduces output, followed by a restriction on the polluting input. Besanko (1987) similarly found that firms profit most from a pollution restriction, though a mandate for an abating input might increase social welfare by increasing output. Thomas's (1980) simulation found the welfare costs of a pollution restriction to be lower than those of other standards, especially those of a mandate on the polluting input. If the results of this study apply to his model as well, then the combination of reduced profits for a firm and reduced output lead to more severe effects by reducing both consumer and producer surplus. Finally, Harford and Karp (1983) focused on comparisons of a pollution-per-output standard versus pollution per input and found the former to be more efficient than the latter,

since it leads to less distortion in input use. The present study, in contrast, found no easily distinguishable differences between these standards (without assumptions on the relative returns to scale of the production and pollution functions): both produce more output than the pollution standard, though both also produce lower profits. The disparity between studies is probably caused by Harford and Karp's assumption of fixed output: while only input ratios could be inefficient in their study, the output level can be inefficient here as well. Given that the results of the other studies are all derived from different models, the similarities in the results outweigh the differences.

The present analysis has also shown that altering the underlying assumptions of the model affects the comparisons of the standards. Any application of this model must consider the specific characteristics of the case to be studied in order to account for the differences caused by different assumptions.

Finally, when a single regulation is imposed on an industry composed of non-identical firms, the chosen standard will not affect all firms identically. The scale, input intensity, and pollution intensity of the firm, among other factors, will determine whether it will satisfy various standards. The chosen standard can thus have an effect on the optimal structure of the firm and may play a role in the entry and exit decisions of firms.

REFERENCES

- Baumol, William J. and Oates, Wallace E., *The Theory of Environmental Policy*, Englewood Cliffs, NJ: Prentice-Hall, 1975.
- Besanko, David, "Performance versus Design Standards in the Regulation of Pollution," *Journal of Public Economics*, October 1987, 34, 19-44.
- Buchanan, James M. and Tullock, Gordon, "Polluters' Profits and Political Response: Direct Controls versus Taxes," *American Economic Review*, March 1975, 65, 139-47.
- Harford, Jon D. and Karp, Gordon, "The Effects and Efficiencies of Different Pollution Standards," *Eastern Economics Journal*, April-June 1983, 9, 79-89.
- Helfand, Gloria E., "Standards versus Standards: The Incentive and Efficiency Effects of Pollution Control Restrictions," Ph.D. Dissertation, Department of Agricultural and Resource Economics, University of California at Berkeley, 1988.
- Hochman, Eithan and Zilberman, David, "Examination of Environmental Policies Using Production and Pollution Microparameter Distributions," *Econometrica*, July 1978, 46, 739-60.
- Magat, Wesley A., Krupnick, Alan J., and Harrington, Winston, *Rules in the Making: A Statistical Analysis of Regulatory Agency Behavior*, Washington, DC: Resources for the Future, 1986.
- Rose-Ackerman, Susan, "Effluent Charges: A Critique," *Canadian Journal of Economics*, November 1973, 6, 512-28.
- Russell, Clifford S., Harrington, Winston, and Vaughan, William J., *Enforcing Pollution Control Laws*, Washington, DC: Resources for the Future, 1986.
- Thomas, Vinod, "Welfare Cost of Pollution Control," *Journal of Environmental Economics and Management*, June 1980, 7, 90-102.

Willingness To Pay and Willingness To Accept: How Much Can They Differ?

By W. MICHAEL HANEMANN*

In many empirical studies, analysts seek to obtain money measures of welfare changes due not to price changes but to changes in the availability of public goods or amenities, changes in the qualities of commodities, or changes in the fixed quantities of rationed goods. The conventional welfare measures for price changes are the compensating (*C*) and equivalent (*E*) variations, which correspond to the maximum amount an individual would be willing to pay (WTP) to secure the change or the minimum amount she would be willing to accept (WTA) to forgo it. Karl-Göran Mäler (1974) was perhaps the first to show that the concepts of *C* and *E* can readily be extended from conventional price changes to such quantity changes. For price changes, Robert Willig (1976) demonstrated that *C* and *E* are likely to be fairly close in value, with the difference depending directly on the size of the income elasticity of demand for the commodity whose price changes. Subsequently, Alan Randall and John Stoll (1980) examined the duality theory associated with fixed quantities in the utility function and showed that, with appropriate modifications, Willig's formulas for bounds on *C* and *E* do, indeed, carry over to this setting.

Within the environmental-economics literature, Randall and Stoll's results have been widely interpreted as implying that WTP and WTA for changes in environmental amenities should not differ greatly unless there are unusual income effects.¹ However,

recent empirical work using various types of interview procedures has produced some evidence of large disparities between WTP and WTA measures.² This has led to something of an impasse: how can the empirical evidence of significant differences between WTP and WTA be reconciled with the theoretical analysis suggesting that such differences are unlikely? Can they be explained entirely by unusual income effects or by peculiarities of the interview process?

In this note, I reexamine Randall and Stoll's analysis and show that, while it is indeed accurate, its implications have been misunderstood. For quantity changes, there is no presumption that WTP and WTA must be close in value and, unlike price changes, the difference between WTP and WTA depends not only on an income effect but also on a substitution effect. By the latter, I mean the ease with which other privately marketed commodities can be substituted for the given public good or fixed commodity, while maintaining the individual at a constant level of utility. I show that, holding income effects constant, the smaller the substitution effect (i.e., the fewer substitutes available for the public good) the greater the disparity between WTP and WTA. This surely coincides with common intuition. If there are private goods that are readily substitutable for the public good, there ought to be little difference between an individual's WTP and WTA for a change in the public good. However, if the public good has almost no substitutes (e.g., Yosemite National Park, or in a different context, your own life), there is no reason why WTP and WTA could not differ vastly: in the limit, WTP could equal the individual's entire (finite)

*Department of Agricultural and Resource Economics, University of California, Berkeley, CA 94720.

¹This view is expressed by, for example, Myrick Freeman (1979 p. 3), Mark A. Thayer (1981 p. 30), Jack L. Knetsch and J. A. Sinden (1984 p. 508), Robin Gregory (1986 p. 326), Don L. Coursey et al. (1987 p. 678), and most of the participants in a recent symposium on valuing amenity resources edited by George L. Peterson et al. (1988 pp. 104, 129, 138, 152, 168, 230, 238, 259).

²See the summaries in table 3.2 of Ronald G. Cummings et al. (1986) and table 1 of Ann Fisher et al. (1988).

income, while WTA could be infinite. My argument is developed in the following two sections. Section I deals specifically with the two polar cases of perfect substitution and zero substitution between the public good and available private goods. Section II deals with Randall and Stoll's extension of Willig's formulas and shows that their bounds are, in fact, consistent with substantial divergences between WTP and WTA. Section III presents empirical application of these bounds and relates them to Mäler's concept of weak complementarity.

I. Two Polar Cases

The theoretical setup is as follows. An individual has preferences for various conventional market commodities whose consumption is denoted by the vector \mathbf{x} as well as for another commodity whose consumption is denoted by q .³ This could represent the supply of a public good or amenity; it could be an index of the quality of one of the private goods; or it could be a private commodity whose consumption is fixed by a public agency.⁴ The key point is that the individual's consumption of q is fixed exogenously, while she can freely vary her consumption of the \mathbf{x} 's. These preferences are represented by a utility function, $u(\mathbf{x}, q)$, which is continuous and nondecreasing in its arguments (I assume that the \mathbf{x} 's and q are all "goods") and strictly quasiconcave in \mathbf{x} . The individual chooses her consumption by solving

$$(1) \quad \max_{\mathbf{x}} u(\mathbf{x}, q) \text{ subject to } \sum p_i x_i = y$$

taking the level of q as given. This yields a set of ordinary demand functions, $x_i = h^i(\mathbf{p}, q, y)$, $i = 1, \dots, N$, and an indirect utility function, $v(\mathbf{p}, q, y) \equiv u[h(\mathbf{p}, q, y), q]$,

³I am treating q as a scalar here, but it could be a vector without seriously affecting the analysis in this section. In the next section, however, the analysis would become significantly more complex if q were a vector and more than one element of q changed.

⁴These alternative interpretations are offered, respectively, by Mäler (1974), Hanemann (1982), and Randall and Stoll (1980).

which has the conventional properties with respect to the price and income arguments and also is nondecreasing in q .⁵ Now suppose that q rises from q^0 to $q^1 > q^0$ while prices and income remain constant at (\mathbf{p}, y) . Accordingly, the individual's utility changes from $u^0 \equiv v(\mathbf{p}, q^0, y)$ to $u^1 \equiv v(\mathbf{p}, q^1, y) \geq u^0$. Following Mäler, the compensating and equivalent variation measures of this change are defined, respectively, by⁶

$$(2) \quad v(\mathbf{p}, q^1, y - C) = v(\mathbf{p}, q^0, y)$$

$$(3) \quad v(\mathbf{p}, q^1, y) = v(\mathbf{p}, q^0, y + E).$$

Dual to the utility maximization in (1) is an expenditure minimization: minimize $\sum p_i x_i$ with respect to \mathbf{x} subject to $u = u(\mathbf{x}, q)$, which yields a set of compensated demand functions, $x_i = g^i(\mathbf{p}, q, u)$, $i = 1, \dots, N$, and an expenditure function, $m(\mathbf{p}, q, u) \equiv \sum p_i g^i(\mathbf{p}, q, u)$, which has the conventional properties with respect to (\mathbf{p}, u) and is non-increasing in q . In terms of this function, C and E are given by

$$(2') \quad C = m(\mathbf{p}, q^0, u^0) - m(\mathbf{p}, q^1, u^0)$$

$$(3') \quad E = m(\mathbf{p}, q^0, u^1) - m(\mathbf{p}, q^1, u^1).$$

It is evident from (2) and (3) that $0 < C < y$ while $E \geq 0$.⁷ The questions at issue are: i) is it true that $E/C \approx 1$? and ii) what factors affect this ratio? As a first cut at an answer, I compare two polar cases. In the first case, at least one private good—say, the first—is a perfect substitute for some

⁵These properties are established in my earlier paper (Hanemann, 1982).

⁶I have taken the liberty of defining C and E as the negative of quantities appearing in Willig (1976) and in Randall and Stoll (1980), so that $\text{sign}(C) = \text{sign}(E) = \text{sign}(u^1 - u^0)$.

⁷I assume throughout that $q^1 > q^0$ and $u^1 \geq u^0$. The analysis could be repeated for a case in which quality decreases and $u^1 < u^0$. In that case, C and E are both nonpositive and correspond, respectively, to the compensation that the individual would be willing to accept to consent to the change and the amount that she would be willing to pay to avoid the change. This would reverse the inequalities presented in what follows, but it would not affect the substance of my argument.

transformation of q . Thus, the direct utility function assumes the special form

$$(4) \quad u(\mathbf{x}, q) = \bar{u}[x_1 + \psi(q), x_2, \dots, x_N]$$

where $\psi(\cdot)$ is an increasing function and $\bar{u}(\cdot)$ is a continuous, increasing, strictly quasi-concave function of N variables. As W. M. Gorman (1976) has shown, assuming an interior solution, the resulting indirect utility function is

$$(5) \quad v(\mathbf{p}, q, y) = \bar{v}[p_1, p_2, \dots, p_N, y + p_1 \cdot \psi(q)]$$

where $\bar{v}(\cdot)$ is the indirect utility function corresponding to $\bar{u}(\cdot)$. Substitution of (5) into (2) and (3) yields the following.⁸

PROPOSITION 1: *If at least one private market good is a perfect substitute for q , then $C = E$.*

At the opposite extreme, I assume that there is a zero elasticity of substitution not just between q and x_1 but between q and *all* the x 's. Thus, the direct utility function becomes

$$(6) \quad u(\mathbf{x}, q) = \bar{u}\left[\min\left(q, \frac{x_1}{\alpha_1}\right), \dots, \min\left(q, \frac{x_N}{\alpha_N}\right)\right]$$

where $\alpha_1, \dots, \alpha_N$ are positive constants and $\bar{u}(\cdot)$ is conventional direct utility function. In this case, the indirect utility function $v(\mathbf{p}, q, y)$ has a rather complex structure and changes its form in different segments of (\mathbf{p}, q, y) -space. It will be sufficient for my purposes to focus on just one of these segments. Suppose that $q \leq y / \sum p_i \alpha_i$; then, the maximization of (6), subject to the budget constraint, yields demand functions and an indirect utility function of the form $x_i = h^i(\mathbf{p}, q, y) = \alpha_i q$, and $u =$

$v(\mathbf{p}, q, y) = \bar{u}(q, \dots, q) \equiv w(q)$. In this region of (\mathbf{p}, q, y) -space, the individual does not exhaust her budget, and her marginal utility of income is therefore zero. Now suppose that $q^0 \leq y / \sum p_i \alpha_i$ and $q^1 > q^0$. Since $v(\mathbf{p}, q^1, y) > w(q^0)$, it is evident from (2) that the individual would be willing to pay some positive but limited amount C to secure this change. However, for any positive quantity E , no matter how large, $v(\mathbf{p}, q^0, y + E) = v(\mathbf{p}, q^0, y) = w(q^0)$. This implies the following proposition.

PROPOSITION 2: *If there is zero substitutability between q and each of the private-market goods, it can happen that, while the individual would only be willing to pay a finite amount for an increase in q , there is no finite compensation that she would accept to forgo this increase.*

It should be emphasized that this result obtains only in a portion of (\mathbf{p}, q, y) space; in other regions, even with (6), E would be finite.⁹ However, the result in Proposition 2 can also be established for other utility functions that permit some substitutability between q and the x 's as long as the indifference curves between q and each of the x 's become parallel to the x axis at some point. The implication of these two propositions is that the degree of substitutability between q and private-market goods *does* significantly affect the relation between C and E . In the next section, I show how this observation can be reconciled with the bounds on C and E derived by Randall and Stoll.

II. Randall and Stoll's Bounds

In order to extend Willig's bounds from price to commodity space, Randall and Stoll focus on a set of demand functions different from those considered above. Suppose that the individual could purchase q in a market

⁸This result carries over, of course, if *more* than one private good is a perfect substitute for q . In the most general case, $u(\mathbf{x}, q) = \bar{u}[x_1 + \psi_1(q), \dots, x_N + \psi_N(q)]$ and $C = E = \sum p_i [\psi_i(q^1) - \psi_i(q^0)]$.

⁹Indeed, if $\bar{h}^i(\alpha_1 p_1, \dots, \alpha_N p_N, y) \leq q^0$, $i = 1, \dots, N$, it can be shown that $v(\mathbf{p}, q^0, y) = v(\mathbf{p}, q^1, y) = \bar{v}(\alpha_1 p_1, \dots, \alpha_N p_N, y)$ and $C = E = 0$, where $\bar{h}^i(\cdot)$ and $\bar{v}(\cdot)$ are the ordinary demand functions and the indirect utility function associated with $\bar{u}(\cdot)$.

at some given price, π . It must be emphasized that this market is entirely hypothetical since q is actually a public good. Instead of (1), she would now solve¹⁰

$$(7) \max_{\mathbf{x}, q} u(\mathbf{x}, q)$$

subject to $\sum p_i x_i + \pi q = y$.

Denote the resulting ordinary demand functions by $x_i = \hat{h}^i(\mathbf{p}, \pi, y)$, $i = 1, \dots, N$ and $q = \hat{h}^q(\mathbf{p}, \pi, y)$. The corresponding indirect utility function is $\hat{v}(\mathbf{p}, \pi, y) \equiv u[\hat{h}(\mathbf{p}, \pi, y), \hat{h}^q(\mathbf{p}, \pi, y)]$. The dual to (7) is: minimize $\sum p_i x_i + \pi q$ with respect to \mathbf{x} and q subject to $u = u(\mathbf{x}, q)$. This generates a set of compensated demand functions, $x_i = \hat{g}^i(\mathbf{p}, \pi, u)$, $i = 1, \dots, N$ and $q = \hat{g}^q(\mathbf{p}, \pi, u)$, and an expenditure function, $\hat{m}(\mathbf{p}, \pi, u) \equiv \sum p_i \hat{g}^i(\mathbf{p}, \pi, u) + \pi \hat{g}^q(\mathbf{p}, \pi, u)$. These functions are hypothetical, since q is really exogenous to the individual, but they are of theoretical interest because they shed light on the relation between C and E .

For any given values of q , \mathbf{p} , and u , the equation

$$(8) \quad q = \hat{g}^q(\mathbf{p}, \pi, u)$$

may be solved to obtain $\pi = \hat{\pi}(\mathbf{p}, q, u)$, the inverse compensated demand (i.e., willingness-to-pay) function for q : $\hat{\pi}(\cdot)$ is the price that would induce the individual to purchase q units of the public good in order to attain a utility level of u , given that she could buy private goods at prices \mathbf{p} . Let $\pi^0 \equiv \hat{\pi}(\mathbf{p}, q^0, u^0)$ and $\pi^1 \equiv \hat{\pi}(\mathbf{p}, q^1, u^1)$ denote the prices that would have supported q^0 and q^1 , respectively. The two expenditure functions dual to (1) and (7) are related by

$$(9) \quad m(\mathbf{p}, q, u) \equiv \hat{m}[\mathbf{p}, \hat{\pi}(\mathbf{p}, q, u), u] \\ - \hat{\pi}(\mathbf{p}, q, u) \cdot q.$$

¹⁰ It is now necessary to assume that $u(\cdot)$ is strictly quasi-concave in both \mathbf{x} and q , rather than \mathbf{x} alone. See footnote 22 for an example in which this is a nontrivial restriction.

This implies that¹¹

$$(10) \quad m_q(\mathbf{p}, q, u) = -\hat{\pi}(\mathbf{p}, q, u).$$

Combining (10) with (2') and (3') yields these alternative formulas for C and E , expressed in terms of the willingness-to-pay function:

$$(2'') \quad C = \int_{q^0}^{q^1} \hat{\pi}(\mathbf{p}, q, u^0) dq$$

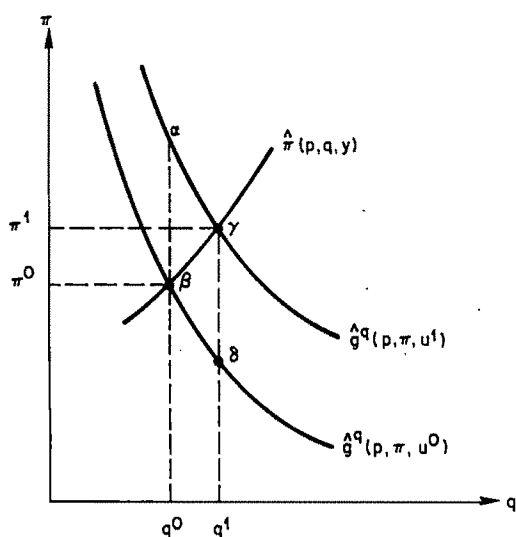
$$(3'') \quad E = \int_{q^0}^{q^1} \hat{\pi}(\mathbf{p}, q, u^1) dq.$$

It can be shown that $\text{sign}(\hat{\pi}_u) = \text{sign}(\hat{h}_y^q)$. Therefore, for given (\mathbf{p}, q) , the graph of $\hat{\pi}(\mathbf{p}, q, u^1)$ lies above (below) that of $\hat{\pi}(\mathbf{p}, q, u^0)$, and $E > (<) C$, when q is a normal (inferior) good. Figure 1 shows E and C for the case in which q is normal: E corresponds to the area $q^0 \alpha \gamma q^1$, while C corresponds to the area $q^0 \beta \delta q^1$.

Using techniques pioneered by Willig (1976), Randall and Stoll (1980) establish bounds on the difference between each of C and E and the area under an inverse ordinary demand function for q . From this, they derive bounds on the difference between C and E . However, the requisite inverse ordinary demand function is obtained in a rather special manner. Given any level of q , one can ask what market price π would induce the individual to purchase that amount of public good if it were available in a market, while still allowing her to purchase the quantity of the x 's that she actually did buy at market prices \mathbf{p} and with income y . In conducting this thought experiment, one needs to supplement the individual's income so that she can afford q as well as the x 's. Thus, for given (\mathbf{p}, q, y) , one seeks the price π that satisfies

$$(11) \quad q = \hat{h}^q(\mathbf{p}, \pi, y + \pi q).$$

¹¹ Using subscripts to denote derivatives, differentiate (9) and note that $q = \hat{g}^q(\mathbf{p}, \pi, u) = \hat{m}_\pi(\mathbf{p}, \pi, u)$ by Shephard's lemma. Equations similar to (9)–(12) are presented by J. P. Neary and K. W. S. Roberts (1980).

FIGURE 1. WTP AND WTA FOR A CHANGE IN q

The solution will be denoted by $\pi = \hat{\pi}(p, q, y)$. This inverse function is related to the inverse compensated demand function by the identities¹²

$$(12a) \quad \hat{\pi}(p, q, y) \\ \equiv \hat{\pi}[p, q, v(p, q, y)]$$

$$(12b) \quad \hat{\pi}(p, q, u) \\ \equiv \hat{\pi}[p, q, m(p, q, u)].$$

Both identities play a role in the analysis. From (12a) it follows that $\pi^0 \equiv \hat{\pi}(p, q^0, u^0) \equiv \hat{\pi}(p, q^0, y)$ and $\pi^1 \equiv \hat{\pi}(p, q^1, u^1) \equiv \hat{\pi}(p, q^1, y)$. Hence, the graph of $\hat{\pi}(p, q, y)$ as a function of q intersects the graph of $\hat{\pi}(p, q, u^0)$ at $q = q^0$, and the graph of $\hat{\pi}(p, q, u^1)$ at $q = q^1$. This is depicted in

¹²Note that $\hat{\pi}(p, q, y)$ is *not* an inverse ordinary demand function in the sense of Ronald W. Anderson (1980), because it involves an income adjustment as well as a price effect.

Figure 1.¹³ From (12b) and (10), one obtains

$$(13) \quad m_q(p, q, u) \\ = -\hat{\pi}[p, q, m(p, q, u)]$$

which is the fundamental differential equation underlying Randall and Stoll's analysis.¹⁴ Also, by differentiating (12b) and then using (13), it can be shown that the concavity of $\hat{m}(p, \pi, u)$ in π , which itself follows from the quasi-concavity of $u(x, q)$ in q , implies the following negativity condition on the Slutsky term associated with $\hat{\pi}(p, q, y)$: $\hat{\pi}_q - \pi \hat{\pi}_y \leq 0$.¹⁵

Define the quantity

$$(14) \quad A \equiv \int_{q^0}^{q^1} \hat{\pi}(p, q, y) dq$$

which corresponds to the area $q^0\beta\gamma\delta q^1$ in Figure 1. This is a sort of Marshallian consumer's surplus, which is to be compared with C and E .¹⁶ Let

$$\xi \equiv \frac{\partial \ln \hat{\pi}(p, q, y)}{\partial \ln y}$$

be the income elasticity of $\hat{\pi}(p, q, y)$; Randall and Stoll call this the "price flexibility of income." Assume that, over the range from (p, q^0, y) to (p, q^1, y) , this elasticity is bounded from below by ξ^L and from above by ξ^U , with neither bound equal to 1. Using

¹³It is commonly supposed the $\hat{\pi}_q < 0$, so that $\pi^0 > \pi^1$ when $q^0 < q^1$ (see e.g., Richard E. Just et al., 1982 fig. 7.12), but this is not correct. It can be shown that $\pi^0 \gtrless \pi^1$ according to whether $\eta \gtrless 1/\alpha$, where η and α are defined in the text below equation (16'). Since $\sum \alpha_i \eta_i + \alpha \eta = 1$ by the Engel aggregation condition, where $\alpha_i \equiv p_i \hat{h}^i(p, \pi, y + \pi q) / (y + \pi q)$ and $\eta_i \equiv (y + \pi q) \hat{h}_y^i(p, \pi, y + \pi q) / \hat{h}(p, \pi, y + \pi q)$, it follows that $\pi^0 \leq \pi^1$ if and only if $\sum \alpha_i \eta_i \leq 0$.

¹⁴This corresponds to Randall and Stoll's equation 7.

¹⁵Also, because $\hat{m}(\cdot)$ is linearly homogeneous in (p, π) , it follows that $\hat{\pi}(p, q, y)$ is linearly homogeneous in (p, y) .

¹⁶Its relation to the conventional Marshallian consumer's surplus associated with the demand for the x 's is analyzed in Proposition 4.

the mean-value theorem, as in Willig's (1976) equation 18, and integrating (13) yields Randall and Stoll's result, namely, the following.

PROPOSITION 3: Assume $\xi^L \leq \xi \leq \xi^U$ where $\xi^L \neq 1$ and $\xi^U \neq 1$. Then,

$$(i) \quad 0 \leq \left[1 + (1 - \xi^L) \frac{A}{y} \right] \frac{1}{1 - \xi^L} - 1 \leq \frac{E}{y}$$

$$(ii) \quad 0 \leq 1 - \left[1 - (1 - \xi^U) \frac{A}{y} \right] \frac{1}{1 - \xi^U} \leq \frac{C}{y} \leq 1.$$

If $\xi^U < 1$, or if $\xi^U > 1$ and $1 + (1 - \xi^U) \frac{A}{y} > 0$,

$$(iii) \quad \frac{E}{y} \leq \left[1 + (1 - \xi^U) \frac{A}{y} \right] \frac{1}{1 - \xi^U} - 1.$$

If $\xi^L > 1$, or if $\xi^L \leq 1$ and $1 - (1 - \xi^L) \frac{A}{y} \geq 0$,

$$(iv) \quad \frac{C}{y} \leq 1 - \left[1 - (1 - \xi^L) \frac{A}{y} \right] \frac{1}{1 - \xi^L}.$$

Applying a Taylor approximation, as in Willig (1976), and assuming that the conditions in (iii) and (iv) are satisfied, one obtains

$$(15) \quad \xi^L \frac{A^2}{2y} \leq E - C \leq \xi^U \frac{A^2}{2y}.$$

This is commonly interpreted as implying that C and E are likely to be close in value, but whether or not that is correct clearly depends on the magnitudes of A/y and the bounds ξ^L and ξ^U . The magnitude of A/y depends in part on the size of the change from q^0 to q^1 ; but what can be said about the likely magnitude of the income elasticity, ξ ? Could it happen, for example, that $\xi^L = \infty$? To answer that question, dif-

ferentiate (11) implicitly:

$$(16) \quad \frac{\partial \hat{\pi}(\mathbf{p}, q, y)}{\partial y} = - \frac{\hat{h}_y^q(\mathbf{p}, \pi, y + \pi q)}{\hat{h}_\pi^q(\mathbf{p}, \pi, y + \pi q) + q \hat{h}_y^q(\mathbf{p}, \pi, y + \pi q)}.$$

By the Hicks-Slutsky decomposition, the denominator is equal to the own-price derivative of the compensated demand function for q and is nonpositive:

$$\begin{aligned} \hat{g}_\pi^q[\mathbf{p}, \pi, v(\mathbf{p}, q, y)] \\ = \hat{h}_\pi^q(\mathbf{p}, \pi, y + \pi q) + q \hat{h}_y^q(\mathbf{p}, \pi, y + \pi q) \\ \leq 0. \end{aligned}$$

Converted to elasticity form, (16) becomes

$$(16') \quad \xi = - \frac{\eta(1 - \alpha)}{\varepsilon}$$

where

$$\eta \equiv \frac{(y + \pi q) \hat{h}_y^q(\mathbf{p}, \pi, y + \pi q)}{\hat{h}_\pi^q(\mathbf{p}, \pi, y + \pi q)}$$

is the income elasticity of the *direct* ordinary demand function for q ,

$$\alpha \equiv \frac{\pi \hat{h}_\pi^q(\mathbf{p}, \pi, y + \pi q)}{y + \pi q}$$

is the budget share of q in relation to "adjusted" income, and

$$\varepsilon \equiv \frac{\pi \hat{g}_\pi^q[\mathbf{p}, \pi, v(\mathbf{p}, q, y)]}{\hat{g}_y^q[\mathbf{p}, \pi, v(\mathbf{p}, q, y)]}$$

is the own-price elasticity of the compensated demand function for q .

The denominator in (16') can be related to the overall elasticity of substitution be-

tween q and the private market goods x_1, \dots, x_N . Assume that the prices p_1, \dots, p_N vary in strict proportion (i.e., $p_i = \theta \bar{p}_i$ for some fixed vector $\bar{\mathbf{p}}$ and some positive scalar θ). Let the aggregate Allen-Uzawa elasticity of substitution between q and the Hicksian composite commodity $x_0 \equiv \sum \bar{p}_i x_i$ be denoted by σ_0 . By adapting W. E. Diewert's (1974) analysis, the following formula can be established relating σ_0 to the compensated own-price elasticity for q : $\varepsilon = -\sigma_0(1 - \alpha)$.¹⁷ Hence, (16') may be written

$$(17) \quad \xi = \frac{\eta}{\sigma_0}.$$

This equation is my fundamental result. It explains the findings in the preceding section about the importance of substitution elasticities. It demonstrates that for changes in q , unlike for changes in p , the extent of the difference between C and E depends not only on income effects (i.e., η) but also on substitution effects (i.e., σ_0). If, over the relevant range, either $\eta = 0$ (no income effects) or $\sigma_0 = \infty$ (perfect substitution between q and one or more of the x 's), then $\xi^L = \xi^U = 0$ and, from Proposition 3, $C = A = E$. On the other hand, if either the demand function for q is highly income elastic or there are very few substitutes for q among the x 's so that σ_0 is close to zero, this could generate very large values of ξ and a substantial divergence between C and E .

¹⁷In deriving this result, one evaluates $\hat{m}(\mathbf{p}, \pi, u)$, $\hat{g}^q(\mathbf{p}, \pi, u)$, and $\hat{g}(\mathbf{p}, \pi, u)$ at $u = v(\mathbf{p}, q, y)$. Hence, the budget shares introduced above satisfy $\alpha \equiv \pi \hat{g}^q(\mathbf{p}, \pi, y + \pi q) / (y + \pi q) = \pi \hat{g}^q(\mathbf{p}, \pi, u) / \hat{m}(\mathbf{p}, \pi, u)$, and $\alpha_j \equiv p_j \hat{h}^j(\mathbf{p}, \pi, y + \pi q) / (y + \pi q) = p_j \hat{g}^j(\mathbf{p}, \pi, u) / \hat{m}(\mathbf{p}, \pi, u)$. In addition to the compensated own-price demand elasticity, ε , I introduce the compensated cross-price demand elasticities $\varepsilon_j \equiv \partial \ln \hat{g}^q(\mathbf{p}, \pi, u) / \partial p_j$. The homogeneity of $\hat{g}^q(\cdot)$ in (\mathbf{p}, π) implies that $\varepsilon + \sum \varepsilon_j = 0$. Under the assumption that $p_j = \theta \bar{p}_j$, $j = 1, \dots, N$, the Allen-Uzawa elasticity of substitution between q and the composite good x_0 is given by $\sigma_0 \equiv [(\partial \hat{g}^q / \partial \theta) \cdot \theta / q] / \alpha_0$ where $\alpha_0 \equiv \sum \alpha_j = 1 - \alpha$ is the expenditure share of x_0 and θ is treated as its price. Observe that $\partial \hat{g}^q / \partial \theta = \sum (\partial \hat{g}^q / \partial p_j) \cdot \bar{p}_j = \sum (\partial \hat{g}^q / \partial p_j) \cdot (p_j / \theta) = (q / \theta) \sum \varepsilon_j = -(q / \theta) \varepsilon$. Hence, $\sigma_0(1 - \alpha) = (\theta / q)(\partial \hat{g}^q / \partial \theta) = -\varepsilon$.

III. Applications

In the first application of (17), the price flexibility of income, ξ , is assumed to be a constant. In that case, as Randall and Stoll note, the bounds in Proposition 3 hold as equalities, and (for simplicity I focus on the case where $\xi \neq 1$):

$$(18a) \quad \frac{C}{y} = 1 - \left[1 - (1 - \xi) \frac{A}{y} \right]^{\frac{1}{1 - \xi}}$$

$$(18b) \quad \frac{E}{y} = \left[1 + (1 - \xi) \frac{A}{y} \right]^{\frac{1}{1 - \xi}} - 1.$$

However, before these formulas can be used to calculate C and E , one must determine how A varies with ξ .

If the price flexibility of income is to be constant, the inverse ordinary demand function must take the form

$$(18c) \quad \hat{\pi}(\mathbf{p}, q, y) = \psi(\mathbf{p}, q) y^\xi$$

where $\psi \geq 0$. Define $G(\mathbf{p}, q) \equiv \int \psi(\mathbf{p}, q) dq$; from (14):¹⁸

$$(18d) \quad A = y^\xi [G(\mathbf{p}, q^1) - G(\mathbf{p}, q^0)].$$

Substituting (18c) into (13) and integrating yields the indirect utility function that generates (18c):

$$(18e) \quad v(\mathbf{p}, q, y)$$

$$= T \left(\left[y^{1 - \xi} + (1 - \xi) G(\mathbf{p}, q) \right]^{\frac{1}{1 - \xi}}, \mathbf{p} \right)$$

where $T(\cdot)$ is some function that is homogeneous of degree zero, increasing in its first argument, and nonincreasing in its other arguments.

¹⁸In order to satisfy the negativity and homogeneity conditions, $\psi(\cdot)$ should be increasing in q and homogeneous of degree $1 - \xi$ in \mathbf{p} . It follows that $G(\cdot)$, too, is homogeneous of degree $1 - \xi$ in \mathbf{p} .

The corresponding demand function for q , $\hat{h}^q(\mathbf{p}, \pi, y)$ can be derived from (18e) by solving $\max_q [v(\mathbf{p}, q, y - \pi q)]$. In general, a closed-form solution cannot be obtained. However, implicit differentiation of the first-order condition for this maximization yields the following expression for the income elasticity of demand:

$$\eta = \frac{\xi \psi + \xi \psi^2 q y^{\xi-1}}{-q \psi_q + \xi \psi^2 q y^{\xi-1}}.$$

It follows that having a constant ξ is generally *not* consistent with having a constant η or a constant σ_0 . An exception occurs when

$$(19a) \quad \xi \psi \equiv -q \psi_q$$

in which case $\eta \equiv 1$ and, from (17), $\sigma_0 \equiv 1/\xi$ (i.e., the price flexibility of income is merely the reciprocal of the elasticity of substitution between q and the x 's). Integrating (19a) yields $\psi(\mathbf{p}, q) = K(\mathbf{p})q^{-\xi}$ for some function $K(\mathbf{p}) \geq 0$ which is homogeneous of degree $1-\xi$, and $G(\mathbf{p}, q) = K(\mathbf{p})q^{1-\xi}/(1-\xi)$. Hence,

$$(19b) \quad A/y = K(\mathbf{p})y^{\xi-1}[(q^1)^{1-\xi} - (q^0)^{1-\xi}]/(1-\xi)$$

and the formulas for C and E become

$$(19c) \quad C/y = 1 - \left\{ 1 - K(\mathbf{p})y^{\xi-1}[(q^1)^{1-\xi} - (q^0)^{1-\xi}] \right\}^{\frac{1}{1-\xi}}$$

$$(19d) \quad E/y = \left\{ 1 + K(\mathbf{p})y^{\xi-1}[(q^1)^{1-\xi} - (q^0)^{1-\xi}] \right\}^{\frac{1}{1-\xi}} - 1.$$

This model with a unitary income elasticity, η , is the only case in which ξ , η , and σ_0 can all be constant simultaneously. It can be shown to be a generalization of the CES utility model $u(\mathbf{x}, q) =$

$[\phi(\mathbf{x})^{1-\xi} + aq^{1-\xi}]^{1/(1-\xi)}$ with homogeneous aggregator function, $\phi(\cdot)$.¹⁹

Equation (19b) makes explicit the dependence of A on ξ . From (19c) and (19d), it follows that the ratio E/C is increasing in both $K(\mathbf{p})$ and ξ . Table 1 tabulates this ratio for several values of K and ξ , for cases where $q^0 = 1$ and $q^1 = 3$. Observe that a low elasticity of substitution ($\sigma_0 \approx 0.07$) can generate a fivefold difference between C and E , even when A/y is very small.²⁰ A similar divergence between C and E can be obtained with a relatively moderate elasticity of substitution ($\sigma_0 = 0.99$), provided that the change matters a lot, in the sense that C/y is large ($C/y \approx 0.8$). However, C is almost identical to E when moderate or large elasticities of substitution are combined with low values of C/y .

The second application is the case in which the inverse demand function takes the form

$$(20a) \quad \hat{\pi}(\mathbf{p}, q, y) = \psi(\mathbf{p}, q)e^{\gamma(\mathbf{p})y}$$

and the price flexibility of income is $\xi = \gamma(\mathbf{p}) \cdot y$, for some $\gamma(\cdot) \geq 0$ which is homogeneous of degree -1 in \mathbf{p} , and some $\psi(\cdot) \geq 0$ which is homogeneous of degree 1 in \mathbf{p} . Substituting (20a) into (13) and integrating yields the indirect utility function that generates (20a):

$$(20b) \quad v(\mathbf{p}, q, y) = T \left(-\frac{e^{-\gamma(\mathbf{p})y}}{\gamma(\mathbf{p})} + G(\mathbf{p}, q), \mathbf{p} \right)$$

¹⁹The difference is that the CES model generates an indirect utility function of the form

$$\bar{v}(\mathbf{p}, q, y) = [y^{1-\xi} + K(\mathbf{p})q^{1-\xi}]^{1/(1-\xi)}$$

whereas the indirect utility function associated with (19a)–(19c) is

$$v(\mathbf{p}, q, y) = T[\bar{v}(\mathbf{p}, q, y), \mathbf{p}].$$

²⁰This is the order of magnitude by which WTA exceeds WTP in some of the empirical studies summarized in table 3.2 of Cummings et al. (1986).

TABLE 1—SIMULATIONS OF WTP AND WTA FOR A GENERALIZED CES UTILITY MODEL

ξ	y	$K(p)$	σ_0	A/y	C/y	E/y	E/C
14	1	0.95	0.0714	0.073	0.05	0.259	5.175
1.01	100	1.4	0.99	1.602	0.796	4.026	5.059
0.677	100	8.1	1.481	2.414	0.991	4.975	5.003
0.677	100	0.1	1.481	0.03	0.029	0.03	1.02

where $G \equiv \int \psi(p, q) dq$, and T is some function that is homogeneous of degree zero, increasing in its first argument, and nonincreasing in the other arguments. The corresponding formula for the elasticity of substitution between q and the x 's, expressed as a function of (p, q, y) , can be shown to be

$$(20c) \quad \sigma_0 = \frac{\psi y + q \psi^2 e^{\gamma y}}{-q y \psi_q + \gamma y q \psi^2 e^{\gamma y}}$$

From (20a) and (20b) it follows that

$$(20d) \quad A = e^{\gamma(p)y} [G(p, q^1) - G(p, q^0)]$$

$$(20e) \quad C = \gamma(p)^{-1} \cdot \ln[1 + \gamma(p)A]$$

$$(20f) \quad E = -\gamma(p)^{-1} \cdot \ln[1 - \gamma(p)A].$$

(note that, when $\gamma A > 1$, $E = \infty$). The ratio E/C is clearly increasing in $\gamma(p)$ and A .

In order to proceed further, it is necessary to take a closer look at A . For this purpose, I focus on the special case of (20) in which $\gamma(p) = \gamma/p_N$, $\psi(p, q) = \delta e^{\alpha + \delta q} p_1^{1-\beta} p_N^\beta / (\beta - 1)$, and

$$(21a) \quad v(p, q, y) =$$

$$T \left(-\frac{p_N}{\gamma} e^{-\gamma y / p_N} + \frac{e^{\alpha + \delta q}}{\beta - 1} p_1^{1-\beta} p_N^\beta, p_2, p_3, \dots, p_N \right)$$

with γ and δ as positive constants and with $\beta > 1$. This implies the following log-log de-

mand function for good 1:²¹

$$(21b) \quad \ln x_1 = \ln h^1(p, q, y) \\ = \alpha - \beta \ln(p_1 / p_N) \\ + (\gamma y / p_N) + \delta q.$$

Since $\lim_{p_1 \rightarrow \infty} \partial v(p, q, y) / \partial q = 0$, good 1 is weakly complementary with q in the sense of Mäler, and C and E can be expressed in terms of the area between the compensated demand curves $g^1(p, q^0, u)$ and $g^1(p, q^1, u)$. Furthermore, in this model the quantity A corresponds to the change in the Marshallian consumer's surplus associated with good 1.

$$(21c) \quad A \equiv \int_{q^0}^{q^1} \hat{\pi}(p, q, y) dq \\ = \left(\frac{p_1}{\beta - 1} \right) [h^1(p, q^1, y) - h^1(p, q^0, y)] \\ = \int_{p_1}^{\infty} [h^1(p, q^1, y) - h^1(p, q^0, y)] dp_1.$$

Hence, the formulas for C and E become

$$(21d) \quad \frac{C}{p_N} = \frac{1}{\gamma} \ln \left\{ 1 + \left(\frac{\gamma}{\beta - 1} \right) \left(\frac{p_1}{p_N} \right) \right. \\ \left. \times [h^1(p, q^1, y) - h^1(p, q^0, y)] \right\}$$

$$(21e) \quad \frac{E}{p_N} = \frac{1}{\gamma} \ln \left\{ 1 - \left(\frac{\gamma}{\beta - 1} \right) \left(\frac{p_1}{p_N} \right) \right. \\ \left. \times [h^1(p, q^1, y) - h^1(p, q^0, y)] \right\}.$$

²¹It follows that, in this model, the income elasticity of demand for good 1 is equal to γy , the price flexibility of income.

TABLE 2—SIMULATIONS OF WTP AND WTA FOR A LOG-LOG UTILITY MODEL

ξ	α	δ	γ	A/y	C/y	E/y	E/C
14	-13.58	0.2	0.14	0.0695	0.0486	0.259	5.334
1.01	1.42	0.3	0.0101	0.959	0.6704	3.414	5.092
0.677	2.151	0.3	0.00677	1.427	0.9987	5.004	5.011
0.677	2.28	0.13	0.00677	0.039	0.0385	0.0396	1.027

These are tabulated in Table 2 for several values of α , γ , and δ for the case in which $p_N = 1$, $p_1 = 1.5$, $\beta = 1.125$, $y = 100$, $q^0 = 1$, and $q^1 = 3$.²² The simulations confirm that $\xi = \gamma y$ and A are the key determinants of the ratio E/C . When either ξ or A is large, then $E \gg C$; when ξ and A are both small, then $E \approx C$.

This example is a striking illustration of the power of Proposition 3. From mere inspection of the ordinary demand function for x_1 in (21b) it is hardly obvious that the term $\xi = \gamma y$ should be a key determinant of the relationship between WTP and WTA for a change in q .²³ The example also raises another issue: the possibility that the quantity A , which forms the basis for Proposition 3, may be related to the conventional Marshallian consumer's surplus associated with a private market commodity, x_1 . Under what circumstances does this carry over to other utility models? Could it, in fact, apply to the

first example based on a generalization of the CES utility model?

By way of answer to the first question, the following lemma establishes that weak complementarity is but one of two conditions that must be satisfied if A is to be equated with a change in the Marshallian consumer's surplus:

LEMMA 1: Suppose there is a private market good, say x_1 , with the properties that (a) it is nonessential and weakly complementary with q , and (b)

$$\frac{\partial \hat{\pi}(\mathbf{p}, q, y)}{\partial p_1} = - \frac{\partial h^1(\mathbf{p}, q, y)}{\partial q}.$$

Then,

$$(22) \quad A = \int_{p_1}^{\infty} [h^1(\mathbf{p}, q^1, y) - h^1(\mathbf{p}, q^0, y)] dp_1.$$

PROOF:

Weak complementarity implies $\lim_{p_1 \rightarrow \infty} m_q(\mathbf{p}, q, u) = 0$. Nonessentialness implies $\lim_{p_1 \rightarrow \infty} m(\mathbf{p}, q, u) < \infty$. Hence, by (13), condition (a) implies $\lim_{p_1 \rightarrow \infty} \hat{\pi}(\mathbf{p}, q, y) = 0$. Accordingly, one can express $\hat{\pi}(\cdot)$ as

$$\begin{aligned} \hat{\pi}(\mathbf{p}, q, y) &= - \int_{p_1}^{\infty} \frac{\partial \pi(\mathbf{p}, q, y)}{\partial p_1} dp_1 \\ &= \int_{p_1}^{\infty} \frac{\partial h^1(\mathbf{p}, q, y)}{\partial q} dp_1 \end{aligned}$$

where the second equality follows from condition (b). Invoking the definition of A in (14) and changing the order of integration yields (22).

²²The parameter values in these simulations are chosen to satisfy the inequalities $(\beta - 1)p_N \leq \gamma p_1 h^1(\mathbf{p}, q, y) \leq \beta p_N$, which ensure that the direct utility function implied by (21) is quasi-concave in \mathbf{x} and q . If $u(\mathbf{x}, q)$ were not quasi-concave in q , the formulas in (21d) and (21e), would still be valid, but the Randall-Stoll bounds in Proposition 3 would not apply. That happens with another special case of (20) in which $\gamma(\mathbf{p}) = \gamma/p_N$ but $\psi(\mathbf{p}, q) = \delta p_N e^{\alpha - \beta(p_1/p_N) + \delta q} / \beta$. This generates an ordinary demand function for good 1 that is identical to (21b), except that $\ln(p_1/p_N)$ is replaced by (p_1/p_N) . Also, (21c)–(21e) hold for this model, except that $\beta - 1$ is replaced by β . However, the implicit utility function $u(\mathbf{x}, q)$ can be shown to be quasi-convex in q .

²³Also, it is hardly obvious that from the demand function in (21b) one can recover σ_0 , the elasticity of substitution between q and the x 's (all the x 's, not just x_1). This is obtained by substituting $\gamma(\mathbf{p}) \equiv \gamma/p_N$ and $\psi(\mathbf{p}, q) \equiv \delta e^{\alpha + \delta q} p_1^{1-\beta} p_N^\beta / (\beta - 1)$ into (20c). The corresponding formula for η can be obtained from $\eta = \sigma_0 \xi = \sigma_0 \gamma y / p_N$.

Application of Lemma 1 yields the answer to the second question:

PROPOSITION 4: *Partition the price vector as $\mathbf{p} = (p_1, \mathbf{p}_{(1)})$. Equation (22) holds if and only if $v(\mathbf{p}, q, y)$ can be expressed in the form*

$$(23) \quad v(\mathbf{p}, q, y) = T[G(\mathbf{p}, q), \mathbf{p}_{(1)}y]$$

where $\lim_{p \rightarrow \infty} G(\mathbf{p}, q) = \lim_{p_1 \rightarrow \infty} G_{p_1}(\mathbf{p}, q) = 0$.

PROOF:

Observe that the conditions on the derivatives of $G(\cdot)$ ensure that x_1 is nonessential and weakly complementary with q . The main task is to show that the functional structure in (23) is necessary and sufficient to satisfy condition (b) of Lemma 1. First use (10) and (12a), and then twice differentiate the implicit function $v(\mathbf{p}, q, y) - u = 0$ to obtain

$$\begin{aligned} \frac{\partial \hat{\pi}(\mathbf{p}, q, y)}{\partial p_1} &= -m_{qp_1}[\mathbf{p}, q, v(\mathbf{p}, q, y)] \\ &\quad - m_{qu}[\mathbf{p}, q, v(\mathbf{p}, q, y)] \cdot v_{p_1}(\mathbf{p}, q, y) \\ &= -(v_q v_{yp_1} - v_y v_{qp_1}) / v_y^2. \end{aligned}$$

Comparing this with $\partial h^1(\mathbf{p}, q, y) / \partial q = (v_{p_1} v_{qy} - v_y v_{qp_1})^2 / v_y^2$, it can be seen that condition (b) will be satisfied if and only if $v_q v_{yp_1} = v_{p_1} v_{qy}$. However, this is equivalent to requiring that (v_q / v_{p_1}) be independent of y , which in turn is equivalent to (23).

The log-log utility model in (21a) clearly satisfies the conditions of Proposition 4. The generalized CES utility model in (18) could also meet the conditions of the proposition, provided that p_1 appears only in the first argument of $T(\cdot)$.

Equation (23) expresses a restriction on the marginal rate of substitution between q and the price of a weakly complementary private-market good, p_1 (i.e., that it be independent of income). This condition was first introduced by Willig (1978) in his paper

on hedonic price adjustments for valuing marginal changes in q : his theorem 1 characterized the circumstances under which the marginal value of q equals the derivative with respect to q of the Marshallian consumer's surplus for x_1 , averaged over the number of units of the good consumed. Proposition 4 expresses a similar result using a different and more compact proof. Combining Propositions 3 and 4 provides a way to value *nonmarginal* changes in q by employing the change in Marshallian consumer's surplus to compute A and then using A to bound WTP or WTA. These two propositions, in effect, establish a new link between Willig's two seminal papers.²⁴

IV. Conclusion

A recent assessment of the state of the art of public-good valuation concludes "Received theory establishes that... WTP... should approximately equal... WTA.... In contrast with theoretical axioms which predict small differences between WTP and WTA, results from contingent valuation method applications wherein such measures are derived almost always demonstrate large differences between average WTP and WTA. To date, researchers have been unable to explain in any definitive way the persistently observed differences between WTP and WTA measures" (Cummings et al., 1986 p. 41.)²⁵ This paper

²⁴I am very grateful to a referee for pointing out the connection with Willig's theorem 1. In my notation, Willig's theorem states that (23) is equivalent to the equality $v_q(\mathbf{p}, q, y) / v_{p_1}(\mathbf{p}, q, y) = \hat{\pi}(\mathbf{p}, q, y) / h^1(\mathbf{p}, q, y)$.

²⁵Some of the debates on divergences between WTP and WTA have focused on the concept of loss-aversion, introduced in the economics literature by Daniel Kahneman and Amos Tversky (1979). This is a different phenomenon from that involved in the Randall-Stoll bounds: it concerns the disparity between the WTP to obtain a change from q^0 to $q^0 + \Delta$ (for some $\Delta > 0$) and the WTP to avoid a change from q^0 to $q^0 - \Delta$, which is not the same as the disparity between WTP and WTA for the same change from q^0 to $q^1 = q^0 + \Delta$. However, the loss/gain disparity can be analyzed using the tools developed in this paper. In a separate paper, I have identified the conditions under which it will exceed the disparity between WTP and WTA studied here.

offers an explanation by showing that the theoretical presumption of approximate equality between WTP and WTA is misconceived. This is because, for public goods, the relation between the two welfare measures depends on a substitution effect as well as an income effect. Given that the substitution elasticity appears in the denominator of (17) and that the Engel aggregation condition places some limit on the plausible magnitude of the income elasticity in the numerator, this suggests that the substitution effects could exert a far greater leverage on the relation between WTP and WTA than the income effects. Thus, large empirical divergences between WTP and WTA may be indicative not of some failure in the survey methodology but of a general perception on the part of the individuals surveyed that the private-market goods available in their choice set are, collectively, a rather imperfect substitute for the public good under consideration.

REFERENCES

- Anderson, Ronald W., "Some Theory of Inverse Demand for Applied Demand Analysis," *European Economic Review*, November 1980, 14, 281-90.
- Coursey, Don L., Hovis, John J. and Schulze, William D., "The Disparity Between Willingness to Accept and Willingness to Pay Measures of Value," *Quarterly Journal of Economics*, August 1987, 102, 679-90.
- Cummings, Ronald G., Brookshire, David S. and Schulze, William D., *Valuing Public Goods: An Assessment of the Contingent Valuation Method*, Totowa, NJ: Rowman and Allanheld, 1986.
- Diewert, W. E., "A Note on Aggregation and Elasticities of Substitution," *Canadian Journal of Economics*, February 1974, 7, 12-20.
- Fisher, Ann, McClelland, Gary H. and Schulze, William D., "Measures of Willingness to Pay versus Willingness to Accept: Evidence, Explanations and Potential Reconciliation," in George L. Peterson, B. L. Driver, and Robin Gregory, eds., *Amenity Resource Valuation: Integrating Economics with Other Disciplines*, State College, PA: Venture, 1988, pp. 127-34.
- Freeman, A. Myrick, *The Benefits of Environmental Improvement: Theory and Practice*, Baltimore: Johns Hopkins University Press, 1979.
- Gorman, W. M., "Tricks With Utility Functions," in M. Artis and R. Nobay, eds., *Essays in Economic Analysis*, New York: Cambridge University Press, 1976, pp. 211-43.
- Gregory, Robin, "Interpreting Measures of Economic Loss: Evidence from Contingent Valuation and Experimental Studies," *Journal of Environmental Economics and Management*, December 1986, 13, 325-37.
- Hanemann, W. Michael, "Quality and Demand Analysis," in Gordon C. Rausser, ed., *New Directions in Econometric Modeling and Forecasting in U. S. Agriculture*, Amsterdam: North Holland, 1982, pp. 55-98.
- Just, Richard E., Hueth, Darrell L. and Schmitz, Andrew, *Applied Welfare Economics and Public Policy*, Englewood Cliffs, NJ: Prentice-Hall, 1982.
- Kahneman, Daniel and Tversky, Amos, "Prospect Theory: An Analysis of Decisions Under Risk," *Econometrica*, March 1979, 47, 263-91.
- Knetsch, Jack L. and Sinden, J. A., "Willingness to Pay and Compensation Demanded: Experimental Disparity in Measures of Value," *Quarterly Journal of Economics*, August 1984, 99, 507-21.
- Mäler, Karl-Göran, *Environmental Economics: A Theoretical Inquiry*, Baltimore: Johns Hopkins University Press, 1974.
- Neary, J. P. and Roberts, K. W. S., "Theory of Household Behavior Under Rationing," *European Economic Review*, January 1980, 13, 25-42.
- Peterson, George L., Driver, B. L. and Gregory, Robin, eds., *Amenity Resource Valuation: Integrating Economics with Other Disciplines*, State College, PA: Venture, 1988.
- Randall, Alan and Stoll, John R., "Consumer's Surplus in Commodity Space," *American Economic Review*, June 1980, 71, 449-57.

Thayer, Mark A., "Contingent Valuation Techniques for Assessing Environmental Impacts: Further Evidence," *Journal of Environmental Economics and Management*, March 1981, 8, 27-44.

Willig, Robert, "Consumer's Surplus Without

Apology," *American Economic Review*, September 1976, 66, 589-97.

———, "Incremental Consumer's Surplus and Hedonic Price Adjustment," *Journal of Economic Theory*, April 1978, 17, 227-53.

Strategic Buyers and the Social Cost of Monopoly

By TORE ELLINGSEN*

Competition for a monopoly position with known profitability is often analyzed as a rent-seeking game, in which the potential producers spend resources in order to obtain the monopoly. Because of such expenditures, the social cost of monopoly is commonly argued to equal the sum of "Harberger costs" (i.e., the deadweight loss) and "Tullock costs" (the expected sum of expenditures on rent seeking).¹ Recently, several writers have recognized that buyers do not always passively accept monopoly pricing. On the contrary, buyers often engage in costly rent-defending activities, such as persuading authorities to regulate the price and quality² of the monopolized good. This occurs not only when the product is an intermediary good sold to a few industrial buyers, but also when consumer organizations engage in legal proceedings and political lobbying on behalf of otherwise disparate and uncoordinated buyers.

At first glance, it is tempting to conclude that the social cost must be higher when the buyers act strategically than when they do not. Indeed, this is the position taken by, for example, John T. Wenders (1987 p. 457),

who claims that, if the buyers engage in costly rent-defending activities, "[T]his may lead to a parallel dissipation of consumers' surplus which is additive to the rent seekers dissipation of producers' surplus." However, there are two reasons why buyer expenditures are not simply additive to the Harberger and Tullock costs. First, when buyers are successful, the deadweight loss is reduced. Secondly, knowledge that monopoly pricing may not be feasible will reduce the producers' interest in the monopoly and hence lower their expenditures. The net impact on welfare from costly buyer activities is therefore far from obvious. The contribution of the present paper is to give precise conditions under which rent-defending activities increase (or decrease) social welfare.

The two formal models that have been proposed in the rent-seeking literature, the "lottery" model of Gordon Tullock (1980) and the "perfectly discriminating contest" model of Jack Hirshleifer and John G. Riley (1978) and Arye L. Hillman and Dov Samet (1987), differ mainly in the assumed relationship between the size of expenditures and the probability of winning the contest. In the lottery model, the probability that a given player wins is proportional to the relative size of his expenditures, whereas in the perfectly discriminating contest the player who spends most resources always wins. As pointed out by Hillman and Riley (1989), it would be desirable to have a unified framework in which the impact of expenditures on political decisions is derived, rather than assumed. In the absence of such a model, however, I have not been able to do better than to study the effect of introducing strategic buyers in each of the two polar cases.

The main conclusions of the analysis are the following. (i) Under the assumption of a perfectly discriminating contest, buyer lobbying is always strictly beneficial from a

*Centre for Applied Research, Norwegian School of Economics and Business Administration, N-5035 Bergen-Sandviken, Norway, and Economics of Industry Group, London School of Economics and Political Science, London WC2A 2AE. The work was financially supported by NORAS and NAVF. I thank John Riley for helpful comments on an earlier version of the paper and for informing me about his own related work (with Arye Hillman). Comments and suggestions from Geir Asheim, Steinar Holden, John Moore, Åsa Rosén and an anonymous referee are also gratefully acknowledged. Remaining errors are mine.

¹Of course, the conclusion hinges in the so-called "wastefulness postulate" (i.e., that the expenditures have no social value). However, all results of this paper can easily be modified to account for less socially wasteful expenditures.

²In what follows, I will assume that the quality is given exogenously. Presumably, the main conclusions would hold if quality as well as price were endogenously determined.

social point of view. (ii) In the lottery model, it is more likely to be beneficial if the monopoly profit is small, if the deadweight loss is large, and if there is a large number of potential producers. (iii) Buyer lobbying is more likely to increase welfare if it occurs after one seller is established than if it occurs at the rent-seeking stage. Indeed, a sufficient condition for costly postentry regulation battles to be socially beneficial under the assumptions of the lottery model is that there are at least three potential sellers. These main results are derived in Sections I, II, and III respectively. Section IV contains some final remarks.

I. The Perfectly Discriminating Contest

Whereas the idea that competition for rents is socially costly is an old one, it is only recently that economists have argued that the whole rent may be dissipated in the competitive process and, hence, that the full monopoly profit should be added to the Harberger costs in measuring the social cost of monopoly (see Tullock, 1967; Anne O. Krueger, 1974; Richard A. Posner, 1975). This conclusion has been supported by the specific game-theoretic formulations of Hirshleifer and Riley (1978) and Hillman and Samet (1987). Here, I first repeat the main ideas of their analyses, and then extend the model to allow for strategic buyers.

A. Inactive Buyers

It is assumed that a number of identical potential producers compete for a monopoly position. Being a monopolist, a seller will quote a price of p_m . This price is higher than the constrained surplus-maximizing price, p_r , which is here taken to be the price associated with zero profit.³ The corre-

sponding quantities are q_m and q_r , respectively. A seller's valuation of winning the contest is equal to the monopoly profit, T . The deadweight loss associated with non-competitive pricing is H . These assumptions are summarized in Figure 1.

Let there be n potential competitors for the monopoly position. The monopoly is awarded to the seller submitting the highest bid in a first-price sealed-bid auction in which both winner and the losers are committed to pay their bids. If the winning bid is not unique, a fair random device will determine the winner. For example, this can be thought of as a bribery game in which each seller is unaware of the size of the bribes made by its competitors and in which the largest bribe, if it is unique, wins with probability 1.

Clearly, there is no pure-strategy equilibrium of this game. If seller i believes that the highest competing bid is $b_j < T$, then the best reply is always to make a slightly higher bid than b_j . If on the other hand it is believed that seller j will bid $b_j = T$, the best reply is $b_i = 0$ for all i ; but then $b_j = T$ is not optimal, since the game can be won more cheaply. A similar argument establishes that no seller bids any given strictly positive amount with positive probability: Suppose that seller i submits $b_i > 0$ with positive probability. Then there will be an interval $[b_i - \varepsilon, b_i]$ on which no competitor will want to bid (because the probability of winning rises discontinuously if a bid in this interval is replaced with one that is slightly higher than b_i). Consequently, seller i can reduce b_i and still win the contest with the same probability, contradicting the assumption that the initial situation was an equilibrium. Hence, in equilibrium, active sellers must use continuous mixed strategies, apart possibly from the bid of 0 which may occur with positive probability.⁴

³This assumption is not substantial, but it simplifies the graphic illustration. If transfers are costless, marginal cost pricing is, of course, welfare-maximizing. Note that Figure 1 is drawn in such a way that it is technologically optimal to have a single seller (the average cost function is downward sloping in the relevant interval because of a fixed cost). It is also worth noting that the assumptions impose no restrictions on the relative magnitudes of T and H .

⁴The interpretation of a mixed-strategy equilibrium in terms of randomized actions is chosen for expositional ease only. As has been pointed out by several writers, a more compelling interpretation is in terms of beliefs: a mixed-strategy equilibrium then simply indicates that the players cannot predict their opponents' actions with full certainty.

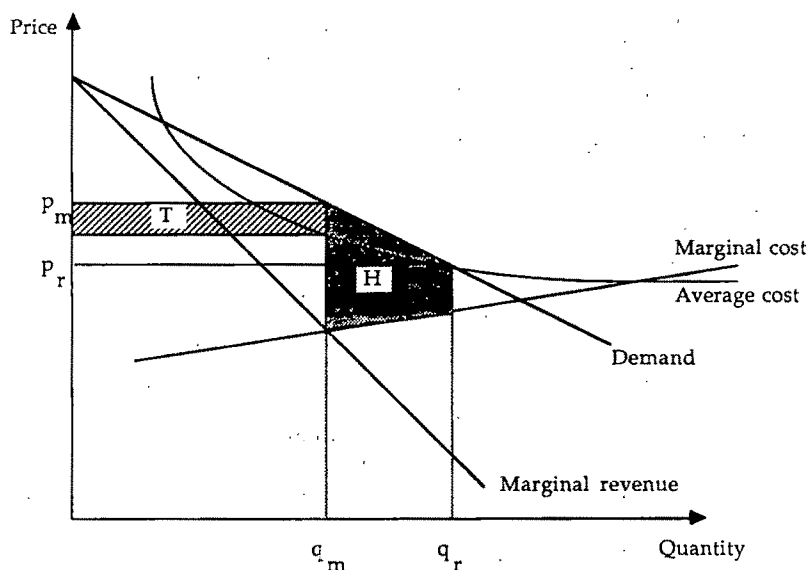


FIGURE 1. PROFIT AND DEADWEIGHT LOSS ASSOCIATED WITH THE UNREGULATED MONOPOLY PRICE.

To see that the equilibrium must involve full rent dissipation (i.e., that total expected rent-seeking expenditures are exactly equal to T), it suffices to demonstrate that the expected profit for each participating seller is 0. Considering only symmetric strategies for sellers that do not bid 0 with probability 1, this is relatively straightforward. Note first that if all players bid 0 with positive probability, the expected profit has to be zero, as a small positive bid would increase the probability of winning discontinuously. Consider now a continuous mixed strategy. Since the lowest bid in the support of this strategy wins with probability 0, the expected profit must be 0 for all bids in the support of this strategy.

Having established that none of the participants expects a strictly positive gain from rent seeking, I arrive at the conclusion that the rent is fully dissipated: the expected sum of rent-seeking expenditures is equal to the monopoly profit. Consequently, the total social cost of monopoly with inactive buyers is $T + H$.

B. Rent-Defending Buyers

What will happen if the buyers refuse passively to accept monopoly pricing in the industry? Suppose buyers are price takers after a certain point in time but they can influence events before this point by taking costly strategic actions. These actions are typically taken to convince some authority to impose a price ceiling or to win direct control over the monopoly supply. I follow Wenders's (1987) example in assuming that, if successful, the buyers will implement the (constrained) surplus-maximizing price p_r .

In this section, the buyer organization is supposed to be strategically symmetric to a potential seller, in that all players take simultaneous moves to gain control over the monopoly. A different and perhaps more realistic assumption is made in Section III.

Whatever actual rent-defending action the buyers take, I will assume that their effectiveness in the contest is the same as that of a single producer. Consequently, the buyers will win the contest if and only if their

outlay is greater than that of the highest-bidding seller. This assumption can of course be relaxed to allow for varying degrees of buyer friendliness by the authorities.

If, as assumed here, all buyers are members of the organization, the value to the organization of winning the contest is equal to the gain in consumers' surplus, $H + T$ (see Ellingsen [1990] for a treatment of the case in which some buyers free ride on the lobbying organization). Note therefore that, in the absence of internal problems, the buyers have a higher valuation of the "prize" than does any of the individual sellers.

Again it is straightforward to check that there is no pure-strategy equilibrium. In particular, it is not an equilibrium to have buyers undertake an expenditure of T , thus securing themselves a gain of H . The reason is that, if the sellers knew that they would have to spend more than T to win, they would all bid 0; but if the sellers bid 0, it is not rational for the buyers to bid as much as T .

In exactly the same way as before, discontinuities of the strategies at other points than $b = 0$ can be ruled out, implying that the equilibrium payoff of sellers is 0. The lower bound of the equilibrium support must also be 0, since the lowest bid is certain not to win (a higher lower bound would imply a negative payoff). Another useful observation is that it never pays for the buyers to bid more than the upper bound of the support of the sellers' mixed strategy (the contest could then be won more cheaply). Furthermore, this upper bound must be equal to T ; a lower upper bound would mean that a seller could earn a positive profit by deviating. From this, it follows that the buyers' equilibrium payoff is H . This is what they earn when bidding T , winning the contest with probability 1, and hence this must be the expected payoff associated with all other bids supporting their mixed strategy.

Now it is easy to see that buyer participation is beneficial. Staying out, the buyers had a payoff of 0. When they participate, they get a payoff of H . In equilibrium, the sellers' payoff is not influenced by the buyers' participation decision. Thus, the social

cost, C , is reduced from $T + H$ to T (i.e., by the full Harberger loss).

For completeness, I give a brief algebraic check on the result. Let $G_S(b)$ be the probability that a particular seller bids less than or equal to b , and define the corresponding probability for the buyers as $G_B(b)$. Suppose that m of the n sellers decide to participate. I consider only equilibria in which the participating sellers behave symmetrically. If the above analysis is correct, the payoff to one of these sellers can then be written as

$$(1) \quad U_S(b) = TG_B(b)[G_S(b)]^{m-1} - b = 0 \\ \text{all } b \in [0, T]$$

and similarly, the buyers' payoff is

$$(2) \quad U_B(b) = (T + H)[G_S(b)]^m - b = H \\ \text{all } b \in [0, T].$$

Solving these two equations for G_S and G_B gives the equilibrium mixed strategies for sellers and buyers, respectively. By construction, all bids in the interval $[0, T]$ give a payoff of 0 to sellers and a payoff of H to buyers. The analysis is completed by checking that no deviation to bids outside the interval yields a higher payoff to any participant. The result is summarized as the following proposition.

PROPOSITION 1: (i) *In the perfectly discriminating contest, there exists a set of n (types of) equilibria, each in which $0 \leq k \leq n - 1$ sellers bid 0 with probability 1, and the remaining $m = n - k$ sellers bid according to the strategy*

$$G_S(b) = \left[\frac{H + b}{T + H} \right]^{1/m}, \quad \text{all } b \in [0, T]$$

and the buyers bid according to

$$G_B(b) = \frac{b}{T} \left[\frac{H + b}{T + H} \right]^{(1-m)/m} \\ \text{all } b \in [0, T].$$

Furthermore, $G_S(b) = G_B(b) = 1$ for all $b > T$. (ii) In this equilibrium, sellers' payoff is 0, buyers' payoff is H , and the total social cost is T , which is less than with inactive buyers.

II. The Lottery Model

The assumption that the biggest spender wins the contest with probability 1 does not seem realistic. An alternative assumption is that the probability of winning is increasing in own expenditure but that the highest bid does not win with certainty.

A model of this kind is proposed by Tullock (1980) and analyzed in detail by Hillman and Riley (1989), the latter paper containing many of the preliminary results derived below. Perhaps surprisingly, the uncertain impact of expenditures affects some of the main conclusions of the previous section.

Below, I will analyze the simplest version of the Tullock model. Here, the probability that participant i will win, P_i , can be written as

$$(3) \quad P_i(b_i) = b_i / \sum_{j=1}^{n+1} b_j$$

where $n+1$ is the number of players. In words, the probability of player i winning is equal to his proportion of the expenditures. Define

$$(4) \quad s_{n+1} \equiv \sum_{j=1}^{n+1} b_j.$$

The expected payoff for player i is then

$$(5) \quad U_i(b_i) = \frac{b_i v_i}{s_{n+1}} - b_i$$

where v_i is player i 's valuation. Since player i 's payoff is no longer discontinuous in the point where b_i is equal to the highest opposing bid, I can now look for an equilibrium in pure strategies. The first-order conditions are easily derived and require that,

if positive, player i 's expenditures are

$$(6) \quad b_i = \frac{s_{n+1} v_i - s_{n+1}^2}{v_i}$$

(remember that s_{n+1} is a function of b_i). It can also be checked that $U_i(\cdot)$ is concave in b_i , so that (6) describes an interior maximum. Assuming that it is profitable to participate for everyone (as it will be), one gets an expression for total expenditures by summing over i in (6):

$$(7) \quad s_{n+1} = \sum_{i=1}^{n+1} \frac{s_{n+1} v_i - s_{n+1}^2}{v_i}$$

and after some trivial manipulations

$$(8) \quad s_{n+1} = \left[\frac{n}{n+1} \right] a_{n+1}$$

where a_{n+1} is defined as

$$(9) \quad a_{n+1} \equiv (n+1) / \sum_{i=1}^{n+1} \frac{1}{v_i}$$

which is the harmonic mean of the players' valuations. Inserting the actual valuations, T for the sellers and $T+H$ for the buyer organization, into (8) one gets

$$(10) \quad s_{n+1} = \left[\frac{n(T+H)}{n(T+H)+T} \right] T.$$

This equation summarizes the relation between rent-seeking expenditures, s_{n+1} , and the number of contesting sellers, n . The rent-seeking expenditures are monotonically increasing in n , never less than half of T , and never exceeding T . Only in the limit, as the number of potential monopolists goes to infinity, is there full rent dissipation. This means that each of the participants has a positive payoff, validating the previous claim that participation is privately profitable [the

local maximum found in eq. (6) is also global].

To determine the social cost, the expected deadweight loss must be added to expected expenditures. The probability that the buyers will win (and that the deadweight loss is 0) is denoted P_B . Similarly, P_S denotes the probability that any given seller will win. Using (8) and (6), equation (3) can then be rewritten as

$$\begin{aligned} (3') \quad P_B &= \frac{n(v_B - v_S) + v_S}{nv_B + v_S} \\ &= \frac{nH + T}{n(T + H) + T} \\ P_S &= \frac{v_S}{nv_B + v_S} = \frac{T}{n(T + H) + T} \end{aligned}$$

Of course, $P_B + nP_S = 1$. Now, let s_n denote the expenditures of the n sellers when buyers are passive, and let a_n be defined accordingly. (Clearly, $a_n = T$.) The net increase in social cost arising from buyer participation is then

$$(11) \quad \Delta C = s_{n+1} - s_n - P_B H.$$

The expression for s_n and s_{n+1} are derived in (8), and P_B is described in (3'). It is therefore straightforward to prove the following.

PROPOSITION 2: *Buyer participation reduces the social cost if and only if $H > T/n$.*

Here, I give the crucial steps of the derivation. From (11), (8), (9), and (3')

$$\begin{aligned} (11') \quad \Delta C &= \frac{n(v_B - v_S) + v_S}{n(nv_B + v_S)} v_S \\ &\quad - \frac{n(v_B - v_S) + v_S}{nv_B + v_S} H. \end{aligned}$$

Inserting $v_B = T + H$ and $v_S = T$, it follows directly that

$$\Delta C < 0 \Leftrightarrow H > T/n.$$

Proposition 2 is rather intuitive. First, when the consumers' surplus is large relative to the profit, the social gain from a buyer victory tends to outweigh the social loss implied by the extra expenditures. Second, when the number of competitors is large, the rent is already almost dissipated by the sellers [see (8)]. Thus, the difference $s_{n+1} - s_n$ is small when n is large. On the other hand, the probability that the buyers will win (and that the efficiency loss is avoided) is always greater than $H/(T + H)$, as can be seen from (3'). Essentially, when n is large, the buyers' expenditures replace, rather than add to, the sellers' expenditures, and the positive probability of eliminating the Harberger costs is the only significant welfare effect.

This result contrasts with that of the previous section in that buyer participation may or may not be socially desirable. The reason is that the degree of rent dissipation in the lottery model depends on the number of competitors. Therefore, contrary to what happened in the perfectly discriminating contest, buyer lobbying strictly reduces the expected payoff for the potential producers. Personally, I find the similarity of the two results more interesting than their differences: expensive buyer lobbying may well increase welfare.

III. More Realistic Timing

So far, I have maintained the assumption that the buyers must compete with all potential sellers in order to win the rent-seeking game. However, in most cases, buyers launch their campaigns for regulation toward existing monopolies, rather than potential ones. Thus, a seller can enter without having to fight the buyers but may face regulation pressure thereafter. Of course, the threat of being regulated influences the desirability of the monopoly and, consequently, the rent-seeking expenditures.

Having established the formal apparatus, it is relatively easy to handle this form of strategic asymmetry. Suppose that the game has two stages: at the first, potential sellers compete for the monopoly position, and at the second, the winner fights the buyer

lobby. I am interested in finding the subgame perfect equilibrium of this two-stage game. Clearly, the second-stage competition is already adequately modeled; one can just set $n = 1$ in the models of Sections I-B and II. The equilibrium strategies of the first stage are then found by letting the sellers compete for the second-stage payoff of the monopolist.

The solution to the two-stage perfectly discriminating contest is particularly easy to find.

PROPOSITION 3: (i) *In the two-stage perfectly discriminating contest there are no rent-seeking expenditures at the first stage. The equilibrium of the second stage is that described in Proposition 1, with $n = 1$.* (ii) *The social cost is T .*

As the expected profit for the monopolist is 0 (Proposition 1, with $n = 1$), no seller is going to spend resources at stage 1, and the monopolist is picked at random. Since all expenditures are made at the second stage, the social cost is the same as with simultaneous moves (i.e., T).

The two-stage version of the lottery model is slightly more difficult to solve. Remember that the social cost consists of three elements: the expenditures incurred at the second stage [found from (10), with $n = 1$], the expenditures incurred at the first stage, and the expected efficiency loss.

The expected second-stage payoff for a seller who has won at the first stage and entered as a monopolist is

$$(12) \quad G \equiv P_S T - b_S.$$

From (3'), when a single seller fights the buyer, $P_S = T/(2T + H)$. The seller's expenditure is found by inserting the expression for total expenditures [eq. (10)], with $n = 1$, into (6). This yields

$$b_S = \frac{n(T + H)T^2}{[n(T + H) + T]^2}$$

and it follows that

$$(12') \quad G = \frac{T^3}{(2T + H)^2}.$$

Obviously, G is the prize for which the sellers will compete at stage 1. From equation (8), the expected expenditures at the first stage can be expressed as $(n - 1)G/n$. As before, the social cost when buyers remain inactive is $(n - 1)T/n + H$. Thus, buyer participation is beneficial if and only if

$$\begin{aligned} \left(\frac{n-1}{n}\right)T + H &> \left(\frac{n-1}{n}\right)\frac{T^3}{(2T+H)^2} \\ &+ \frac{T(T+H)}{2T+H} \\ &+ \left(\frac{T}{2T+H}\right)H. \end{aligned}$$

It is a matter of fairly straightforward manipulations to rewrite this condition as

$$(13) \quad \frac{(T + H)[(2n - 1)HT + (n - 3)T^2 + nH^2]}{n(2T + H)^2} > 0.$$

This yields the following proposition.

PROPOSITION 4: *In the two-stage version of the lottery model, two (different) sufficient conditions for buyer participation to be beneficial are (i) $n \geq 3$ and (ii) $n = 2$ and $H > (\sqrt{17} - 3)T/4 \approx T/3.56$.*

Condition (i) is seen directly from (13), whereas condition (ii) amounts to finding the roots of the equation: $2H^2 + 3HT - T^2 = 0$.

Proposition 4 is quite striking. A sufficient condition for buyer participation to reduce social cost is that there are at least three sellers competing for the monopoly position. Even with only two sellers, the

condition is considerably weaker ($H > T/3.56$) than under the assumption of simultaneous moves ($H > T/2$). Thus, the two-stage formulation lends further support to the claim that buyer lobbying tends to increase welfare.

IV. Final Remarks

The analysis points out three effects of buyer participation in rent-seeking games on the social cost of monopoly: the direct and detrimental effect of one extra player's expenditures, the indirect effect on opponents' expenditures, and the saved consumers' surplus.

In this paper, I have presented conditions under which expensive buyer lobbying is socially beneficial. Overall, these conditions are weak. When the highest bidder certainly wins, voluntary buyer spending is always beneficial. If the impact of spending is uncertain, the impact on welfare of costly buyer activities is ambiguous,⁵ but when buyers only need to fight an incumbent monopolist, lobbying is very likely to be beneficial.

In conclusion, I will briefly relate the above results to the work of other authors. The debt to the work of Hillman and Riley (1989) has already been acknowledged. Their general analysis of contests with asymmetric valuations is ideally suited for tackling the questions addressed here. Pointing out that asymmetric valuations will

lead to smaller expenditures in the perfectly discriminating contest, these authors anticipated the message of my Proposition 1. A similar result is also obtained by Wing Suen (1989).

To my knowledge, the first systematic attempt to highlight the role of strategic buyers in rent seeking is Barry Baysinger and Robert D. Tollison (1980; see also Tollison [1982]). One of their claims is that, when the impact of expenditures is uncertain, the total social cost of monopoly may exceed the sum of deadweight loss and monopoly profit. It should be clear from the above analysis that this is impossible. Essentially, the authors overlook the elementary fact that when expenditures are voluntary, their sum cannot exceed the prize: the buyers can earn at least the consumers' surplus associated with monopoly pricing; the sellers can secure themselves 0. They cannot jointly earn less than this in equilibrium. Hence, expected social cost cannot be larger than $T + H$. More recently, Wenders (1987) has also argued that costly rent-defending activities would unambiguously increase the social cost. Not using game-theoretic tools, he fails to recognize the effect of buyer expenditures on the behavior of other participants.

Finally, the results obtained here can be compared to those of Elie Applebaum and Eliakim Katz (1986). They analyze a model in which a number of interest groups compete for a transfer which is to be collected from the losers. They demonstrate that the participants spend more resources in this environment than if the amount of transfer had first been collected and only then contested. They draw the conclusion that costs incurred to avoid transfers tend to increase the total social cost of transfer seeking. However, this result stems simply from the fact that more is at stake for each participant in the former case than in the latter. The present paper shows that, if there is costly competition for a transfer from some *predesignated* group and there is some loss associated with the transfer, society will often gain if this group spends resources in order to avoid the transfer altogether.

⁵However, it should be mentioned that this ambiguity need not be an inherent feature of all models in which the impact of expenditures is uncertain. The lottery model exhibits some dubious properties. Particularly, entry of a new competitor never induces exit of any of the already participating sellers. In any model in which entry of buyers into the game induces exit of some seller(s), buyer participation seems likely to be beneficial. Hillman and Riley (1989) have argued that Tullock's stochastic specification, as represented by equation (3), is not founded upon any basic principle of how the uncertainty of a dollar's impact should vary with the total amount of money spent by the agent. They also present an example in which the relationship between bids and the probability of winning is derived rather than assumed. In that case, the number of participants in equilibrium is limited.

APPENDIX

Strategic Buyers and the Social Cost of Monopoly: Some Derivations

i) The manipulation leading from (7) to (8) is as follows:

$$s_{n+1} = \sum_{i=1}^{n+1} \frac{s_{n+1}v_i - s_{n+1}^2}{v_i}$$

$$1 = \sum_{i=1}^{n+1} \left[1 - \frac{s_{n+1}}{v_i} \right]$$

$$1 = n+1 - s_{n+1} \sum_{i=1}^{n+1} \frac{1}{v_i}$$

$$s_{n+1} = \frac{n}{\sum_{i=1}^{n+1} \frac{1}{v_i}}$$

which completes the demonstration [substitute the definition of a_{n+1} in (8)].

ii) From the above expression, it is also easy to derive (10):

$$\begin{aligned} s_{n+1} &= \frac{n}{\frac{n}{v_S} + \frac{1}{v_B}} = \frac{nv_B v_S}{nv_B + v_S} \\ &= \left(\frac{n(T+H)}{n(T+H)+T} \right) T \end{aligned}$$

iii) The derivation of (3'):

$$\begin{aligned} P_i &= \frac{b_i}{s_{n+1}} = \frac{\frac{s_{n+1}(v_i - s_{n+1})}{v_i}}{s_{n+1}} \\ &= \frac{v_i - s_{n+1}}{v_i} \end{aligned}$$

and thus,

$$\begin{aligned} P_B &= \frac{T+H - \left(\frac{n(T+H)}{n(T+H)+T} \right) T}{T+H} \\ &= 1 - \frac{nT}{n(T+H)+T} \\ &= \frac{T+nH}{n(T+H)+T} \\ P_S &= \frac{T - \left(\frac{n(T+H)}{n(T+H)+T} \right) T}{T} \\ &= 1 - \frac{n(T+H)}{n(T+H)+T} \\ &= \frac{T}{n(T+H)+T} \end{aligned}$$

iv) The omitted step in the proof of Proposition 2 is from (11) to (11'). Here, I find the expression for $s_{n+1} - s_n$:

$$\begin{aligned} \frac{n}{\frac{n}{v_S} + \frac{1}{v_B}} - \frac{n-1}{\frac{n-1}{v_S}} &= \frac{\frac{n}{v_S}}{\left[\frac{n}{v_S} + \frac{1}{v_B} \right] \frac{n}{v_S}} \\ &\quad - \frac{(n-1) \left[\frac{n}{v_S} + \frac{1}{v_B} \right]}{\left[\frac{n}{v_S} + \frac{1}{v_B} \right] \frac{n}{v_S}} \\ &= \frac{\frac{n}{v_S} - \frac{n-1}{v_B}}{\left[\frac{n}{v_S} + \frac{1}{v_B} \right] \frac{n}{v_S}} \\ &= \frac{nv_B - (n-1)v_S}{n \left[\frac{nv_B}{v_S} + 1 \right]} \\ &= \frac{n(v_B - v_S) + v_S}{n(nv_B + v_S)} v_S \end{aligned}$$

which is the first part of equation (11').

v) In the derivation of (13), the only trick is to show that

$$\begin{aligned} & (n-3)T^3 + (3n-4)HT^2 \\ & + (3n-1)H^2T + nH^3 \\ & = (T+H)[nH(T+H) + nHT \\ & - T(T+H) - 2T^2 + nT^2]. \end{aligned}$$

REFERENCES

- Appelbaum, Elie and Katz, Eliakim, "Transfer Seeking and Avoidance: On the Full Cost of Rent Seeking," *Public Choice*, 1986, 48 (2), 175-81.
- Baysinger, Barry and Tollison, Robert D., "Evaluating the Social Cost of Monopoly and Regulation," *Atlantic Economic Journal*, December 1980, 8, 22-6.
- Ellingsen Tore, "Strategic Buyers and the Social Cost of Monopoly," Discussion Paper, Department of Economics, Norwegian School of Economics and Business Administration, May 1990.
- Hillman, Arye L. and Riley, John G., "Politically Contestable Rents and Transfers," *Economics and Politics*, Spring 1989, 1, 17-39.
- _____, and Samet, Dov, "Dissipation of Rents and Revenues in Small Numbers Contests," *Public Choice*, 1987 54 (1), 63-82.
- Hirshleifer, Jack and Riley, John G., "Auctions and Contests," UCLA Working Paper No. 118B, 1978.
- Krueger, Anne O., "The Political Economy of the Rent-Seeking Society," *American Economic Review*, June 1974, 64, 291-303.
- Posner, Richard A., "The Social Costs of Monopoly and Regulation," *Journal of Political Economy*, August 1975, 83, 807-27.
- Suen, Wing, "Rationing and Rent Dissipation in the Presence of Heterogeneous Individuals," *Journal of Political Economy*, December 1989, 97, 1384-94.
- Tollison, Robert D., "Rent Seeking: A Survey," *Kyklos*, 1982, 35 (4), 575-602.
- Tullock, Gordon, "The Welfare Costs of Tariffs, Monopolies and Theft," *Western Economic Journal*, June 1967, 5, 224-32.
- _____, "Efficient Rent Seeking," in James M. Buchanan, Robert D. Tollison, and Gordon Tullock, eds., *Toward a Theory of the Rent Seeking Society*, College Station: Texas A&M University Press, 1980, 269-82.
- Wenders, John T., "On Perfect Rent Dissipation," *American Economic Review*, June 1987, 77, 457-9.

Increasing the Profits of a Subset of Firms in Oligopoly Models with Strategic Substitutes

By GÉRARD GAUDET AND STEPHEN W. SALANT*

Consider an industry composed of N firms in a symmetric equilibrium. Designate a subset of S ($\leq N$) firms and marginally reduce the strategic variables of the firms in the subset. If the remaining firms simultaneously make the best reply to this exogenous displacement, under what circumstances will profits of the firms in the designated subset increase?

We show that, in the case of Cournot competition among producers of perfect substitutes, a marginal contraction is strictly beneficial (strictly harmful) if and only if the number of firms in the designated subset exceeds the "adjusted" number of firms outside it by strictly more (strictly less) than one. The adjustment factor is unity when cost and demand functions are linear but, more generally, depends on the convexity of the cost and demand curves. Thus, a marginal contraction of two firms in a triopoly has no effect on the profits of firms in the subset if cost and demand functions are linear; if instead cost is linear but the inverse demand function is strictly concave (strictly convex), a marginal contraction will strictly decrease (strictly increase) profits. The analysis is extended to the effects of *nonmarginal* exogenous changes in the outputs of the constrained firms. This extension has implications for the relationship of Stackelberg (sequential move) and Nash (simultaneous move) equilibria.

Our analysis has broad application. To illustrate, we show that it unifies results in the literature on export subsidies, horizontal mergers, and strikes.

The paper is organized as follows. In Section I, our comparative-static result is derived. In Section II, we present a wide range of applications. In the conclusion and the Appendix we generalize our analysis to other situations involving strategic substitutes.

I. The Effect of an Exogenous Output Contraction

A. Marginal Changes

Consider an industry composed of N firms producing a homogeneous good with identical cost functions. Denote the inverse market demand by $p(Q)$. It is assumed that there is a $Q^0 > 0$ such that $p(Q) > 0$ for $Q < Q^0$ and $p(Q) = 0$ for $Q \geq Q^0$ and that $p(Q)$ is twice continuously differentiable with $p'(Q) < 0$ for $Q < Q^0$. The typical firm's cost of production is $C(q)$, with $C(q) > 0$ if $q > 0$ and $C(0) = 0$. It is assumed to possess continuous first and second derivatives, which satisfy $C' \geq 0$ and $C'' \geq 0$.

We further assume that $p'(Q) + qp''(Q) < 0$ for all $0 \leq q \leq Q \leq Q^0$. Thus, a given firm's marginal profit must fall when any rival firm increases its output. The goods are therefore strategic substitutes: each firm's best-reply function is downward sloping.¹

These assumptions insure that a unique symmetric Cournot equilibrium exists in

*Département des Sciences Économiques, Université du Québec à Montréal, and Department of Economics, University of Michigan. We thank two anonymous referees as well as Mark Bagnoli, Marcel Boyer, Andrew Daughety, Avinash Dixit, Abraham Hollander, Jeffrey Perloff, Greg Shaffer, Carl Shapiro, Joe Swierzbinski, and Hal Varian for their comments on an earlier version of this paper. Gérard Gaudet acknowledges financial support from the Social Science Research Council of Canada.

¹See Jeremy I. Bulow et al. (1985) for a general definition of strategic substitutes and complements. We discuss briefly the implications of our analysis for strategic complements in the concluding section and the Appendix.

pure strategies.² Now assume that this initial equilibrium is displaced. In particular, assume that each firm in a designated subset of $S \leq N$ firms is constrained to produce an exogenous amount, \bar{q} . The output of each of the unconstrained $N - S$ firms is endogenous. We denote it by q and, to emphasize its dependence on \bar{q} , sometimes write $q(\bar{q}, N, S)$.

Let firm i be a member of the constrained subset and let π^s denote the profit of this firm. It can be expressed as a function of two variables, own output and the aggregate output of the other $N - 1$ firms:

$$(1) \quad \pi^s(\bar{q}, Q_{-i}) = \bar{q}p(\bar{q} + Q_{-i}) - C(\bar{q})$$

where $Q_{-i} = (N - S)q + (S - 1)\bar{q}$ denotes the aggregate output of every firm other than i .

If $\bar{q} = q$, firm i 's output is a best reply and

$$(2) \quad \left. \frac{\partial \pi^s(\bar{q}, Q_{-i})}{\partial \bar{q}} \right|_{\bar{q}=q} = 0.$$

Totally differentiating (1) and using (2) to evaluate the derivative at $\bar{q} = q$, we obtain

$$(3) \quad \left. \frac{d\pi^s}{d\bar{q}} \right|_{\bar{q}=q} = \frac{\partial \pi^s(\bar{q}, Q_{-i})}{\partial Q_{-i}} \frac{dQ_{-i}}{d\bar{q}} = \bar{q}p' \frac{dQ_{-i}}{d\bar{q}}.$$

Since $p' < 0$, it follows that a marginal change (increase or decrease) in \bar{q} in a neighborhood of the initial equilibrium will raise the profit of any given firm in the constrained subset if and only if the equilibrium aggregate output of the $N - 1$ other firms declines as a consequence. From the definition of Q_{-i} ,

$$(4) \quad \frac{dQ_{-i}}{d\bar{q}} = (N - S) \frac{dq}{d\bar{q}} + (S - 1).$$

It follows that

$$(5) \quad \left. \frac{d\pi^s}{d\bar{q}} \right|_{\bar{q}=q} \gtrless 0 \Leftrightarrow (N - S) \frac{dq}{d\bar{q}} + (S - 1) \gtrless 0.$$

Since the goods are strategic substitutes, each of the $N - S$ unconstrained firms responds to a marginal reduction in \bar{q} by *increasing* its output ($dq/d\bar{q} < 0$). In the extreme case where $S = 1$ and $N > 1$, then $d\pi^s/d\bar{q} > 0$. A marginal contraction in \bar{q} increases the aggregate output of the firms other than i since there are no other firms *inside* the constrained subset; hence, firm i 's profits must fall. This explains why a Stackelberg leader who takes over one firm in a duopoly will instead *expand* its production. If, at the other extreme, $S = N > 1$, then $d\pi^s/d\bar{q} < 0$. In that case, a marginal contraction in the output of the S firms will cause Q_{-i} to fall since there are no firms *outside* the constrained subset; hence, the profits of firm i must rise. This explains why it is always profitable for a monopolist who takes over an industry of N independent firms to contract the production of each. In these polar cases, the qualitative results do not depend on the responsiveness of the outside firms to the exogenous contraction in the outputs of the S firms.

When $1 < S < N$, however, the *magnitude* of the response of the unconstrained firms does determine whether a marginal contraction of the outputs of the S firms increases or decreases their profits. The equilibrium output $q(\bar{q}, N, S)$ of each unconstrained firm solves the following first-order condition:

$$(6) \quad p(S\bar{q} + [N - S]q) + qp'(S\bar{q} + [N - S]q) - C'(q) = 0$$

from which we derive

$$(7) \quad \frac{dq}{d\bar{q}} = \frac{-S\alpha}{(N - S)\alpha + 1} = \frac{-S}{(N - S) + 1/\alpha} < 0$$

²For details, see Gaudet and Salant (1991b).

where

$$(8) \quad \alpha = \frac{p' + qp''}{p' - C''} > 0.$$

The larger is α , the more responsive are the unconstrained firms.³ The magnitude of α in turn depends upon the slope of the marginal cost function of each unconstrained firm as well as the curvature of the demand function. This is intuitively plausible: with linear demand, for example, an increase in the slope of the marginal cost curve (C'') of each unconstrained firm reduces α and hence reduces responsiveness to changes in \bar{q} .

Substituting the expression for this response from (7) into (5), we get

$$(9) \quad \left. \frac{d\pi^s}{d\bar{q}} \right|_{\bar{q}=q} \leq 0 \Leftrightarrow S-1 \geq \alpha(N-S).$$

In the case where $\alpha = 1$ (for example, when $C'' = p'' = 0$) condition (9) says that the firms in the subset will strictly gain from an exogenously induced marginal contraction of their output if and only if they outnumber the firms outside the subset by more than one. Refer to the term $\alpha(N-S)$ as the "adjusted" number of firms outside the subset. Then, more generally, a marginal output contraction is profitable for the firms in the subset if and only if they exceed the adjusted number of firms outside the subset by more than one.⁴

B. Nonmarginal Changes

A geometric illustration of our results will be helpful in analyzing nonmarginal changes. For any exogenous \bar{q} , each uncon-

strained firm will produce $q(\bar{q}, N, S)$, as defined by (6). We represent this in Figure 1 as a downward-sloping locus in (\bar{q}, q) -space with slope given by (7).⁵ For simplicity, we have drawn it as a line. The unconstrained symmetric equilibrium occurs where the upward-sloping 45-degree line intersects this downward-sloping locus.

Denote the aggregate profits of the S firms in the subset by Π^s :

$$(10) \quad \Pi^s(\bar{q}, q) = S[\bar{q}p(S\bar{q} + [N-S]q) - C(\bar{q})].$$

The slope of the "isoprofit curve" through the unconstrained equilibrium point ($\bar{q} = q$) is therefore

$$(11) \quad \left. \frac{d\bar{q}}{dq} \right|_{\bar{q}=q} = -\frac{S-1}{N-S}.$$

Notice that at $\bar{q} = q$ this slope depends only on the numbers (N, S) and not on the curvature of the demand or cost curves. Moreover, it is negative, except in the over-studied special case in which $S = 1$.⁶ Since a reduction in q with \bar{q} fixed would raise price and increase the profits of the S firms, points below this isoprofit curve represent higher profits to the S firms. We will assume that, for any finite N , $\Pi^s(\bar{q}, q(\bar{q}, N, S))$ is a strictly quasi-concave function of \bar{q} .⁷ Our assumptions are jointly sufficient to insure that each isoprofit line cuts the downward-sloping locus at most twice and that $\Pi^s(\bar{q}, q(\bar{q}, N, S))$ has a unique maximum as a function of \bar{q} .

³Farrell and Shapiro (1990) make use of the same expression as α in their analysis of the welfare effects of horizontal mergers.

⁴Notice that α and hence (9) depend on the curvature of the cost functions of the unconstrained firms but not of the constrained firms. Thus, as pointed out by an anonymous referee, the assumption that the constrained and unconstrained firms are identical is not necessary.

⁵Equation (11) can be readily interpreted. If each of the S firms contracts by one unit, then an expansion by each of the $N-S$ firms by $(S-1)/(N-S)$ units would result in a zero net expansion of the $N-1$ firms other than the designated constrained firm and hence would not change the profits of the S constrained firms.

⁶The locus will be linear for any (N, S) when the demand and cost curves are linear. In the most familiar case with $S=1$ and $N=2$, the locus is the linear "reaction function" with slope $-1/2$.

⁷This assumption will be satisfied if, for example, demand is linear and marginal cost is constant. In that case $\Pi^s(\bar{q}, q(\bar{q}, N, S))$ is strictly concave in \bar{q} .

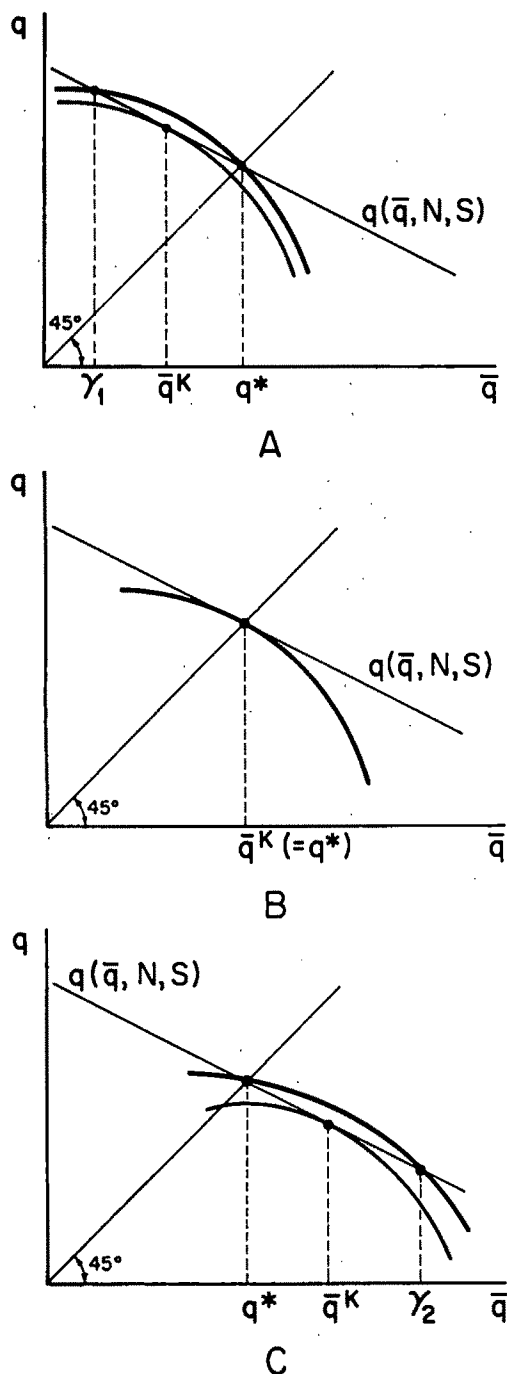


FIGURE 1. NONMARGINAL CHANGES WHEN THE ISOPROFIT CONTOUR THROUGH THE UNCONSTRAINED EQUILIBRIUM POINT IS (A) STEEPER, (B) AS STEEP, OR (C) FLATTER THAN THE DOWNWARD-SLOPING LOCUS DEFINED BY EQUATION (6).

Three possibilities can arise. The isoprofit contour through the unconstrained equilibrium point can be (A) steeper, (B) as steep, or (C) flatter than the downward-sloping locus defined by the first-order condition (6); that is,

$$(12) \quad -\frac{S-1}{N-S} \gtrless \left. \frac{dq(\bar{q}, N, S)}{d\bar{q}} \right|_{\bar{q}=q}$$

where the right-hand side is given by (7) evaluated at $\bar{q} = q$. A marginal contraction in \bar{q} will be beneficial to the S firms in case (A), neutral in case (B), and injurious in case (C). This clarifies (5) geometrically. We illustrate cases (A), (B), and (C) respectively in the three panels of Figure 1.

As we have previously shown, case (A) occurs when $S - \alpha(N - S) > 1$. In that case, a *marginal* contraction would be beneficial. However, note that there is some level of \bar{q} for which the profits of the S firms would be exactly the same as in the unconstrained equilibrium. Denote this level of \bar{q} by γ_1 . Geometrically, γ_1 is the horizontal coordinate of the point where the isoprofit contour through the unconstrained equilibrium point cuts the downward-sloping locus to the *left* of the unconstrained equilibrium point. Analytically, γ_1 is the value of $\gamma < q^*$ that solves⁸

$$(13) \quad p(S\gamma + [N - S]q(\gamma, N, S))\gamma - C(\gamma) \\ = p^*q^* - C(q^*)$$

where p^* and q^* are the price and outputs in the unconstrained Cournot equilibrium. Reducing the output of each constrained firm below the critical level γ_1 will lower their profits relative to the unconstrained equilibrium.

Similarly, case (C) occurs when $S - \alpha(N - S) < 1$. In that case, a *marginal* expansion would be beneficial. However, if the constrained firms were compelled to expand their outputs to γ_2 , their profits would be

⁸Our assumptions insure that (13) yields a unique value for γ .

reduced to the level in the unconstrained equilibrium. Geometrically, γ_2 is the horizontal coordinate of the point where the isoprofit contour through the unconstrained equilibrium point cuts the downward-sloping locus to the *right* of the unconstrained equilibrium point. Analytically, γ_2 is the value of $\gamma > q^*$ that solves equation (13). Increasing the output of each constrained firm above γ_2 will lower profits below their initial level.

II. Applications

The comparative-static results we have been investigating have many applications. Consider first the effect of a labor strike on the profits of the struck firms.⁹ It is often taken for granted that a struck firm must be harmed by the forced reduction in its output.¹⁰ However, our analysis delineates circumstances in which, under an oligopolistic market structure, each targeted firm will in fact benefit from a strike.¹¹ The 1979 strike against selected lettuce producers in the Imperial Valley of California appears to have illustrated this perverse result.¹²

Our results imply that a marginal strike is strictly beneficial to each struck firm if and only if the number of struck firms exceeds the (adjusted) number of nonstruck firms by more than one. Under the same circumstances, a nonmarginal strike will also be beneficial as long as the output of each

struck firm is not forced below the critical level, $\gamma_1 < q^*$, defined in (13).

As a second application, consider the rationale for subsidizing the exports of domestic firms discussed recently by James A. Brander and Barbara J. Spencer (1985).¹³ They consider a model in which Cournot duopolists, one in each country, export to consumers in a third country. They show that imposition of a marginal subsidy, in the neighborhood of free trade equilibrium, will always increase the profit of the domestic firm *net of the subsidy*.¹⁴ These results rely again on the comparative-static properties of the Cournot model derived in Section I. Our analysis makes clear that the Brander and Spencer result is a special case: regardless of the curvatures of the cost and demand curves, if there is only one domestic firm in the subset ($S = 1$) and one or more foreign firms ($N - S \geq 1$), then a marginal expansion of the output of the firm in the subset in the neighborhood of Cournot equilibrium must increase its profits. This clearly follows from (9). In an N -firm oligopoly, however, the domestic attractiveness of the subsidy will depend on the number of domestic firms.¹⁵ In fact, the optimal policy may instead be a tax.

⁹A more timely example might be the effects on the profits of a subset of the world's coca growers if the "war on drugs" succeeds in slightly reducing their production.

¹⁰A strike might not only reduce output but might in addition increase the cost of producing that reduced output. We ignore this additional effect but note that, even in its presence, each targeted firm might still benefit from the strike. A similar point can be made about the war on drugs discussed in the previous footnote.

¹¹Under a competitive market structure, each targeted firm will in fact always benefit from a "marginal" strike, no matter how small a proportion of the firms are targeted. For a formal discussion of the comparative-static results in the competitive case, see Gaudet and Salant (1989).

¹²See Colin A. Carter et al. (1981) for an empirical analysis of the effects of this strike.

¹³The issues raised by Brander and Spencer's model are not confined to the international arena. To the extent that any government, be it city, state, or national, wishes to increase the profits of producers, its concern is restricted to the producers within its own jurisdiction; but those producers constitute only a subset of the industry when some firms lie beyond that jurisdiction. For example, California subsidizes the water used by its farmers. While this affects profits of farms in other states, it is the profits of in-state farms that the California government is concerned to influence.

¹⁴Jonathan Eaton and Gene M. Grossman (1986) also consider Bertrand competition in their analysis of optimal trade policy under oligopoly. They show that, whether the duopolists compete in price or quantity, an export tax (rather than subsidy) is optimal if the goods are strategic complements (upward-sloping reaction functions) but not complements in demand. The export subsidy is optimal if the goods are strategic substitutes. The extension of our analysis to strategic complementarity and demand complementarity, discussed in the concluding section and the Appendix, will clarify the intuition behind this result.

¹⁵On this point, see also Avinash Dixit (1984) and Salant (1984). Referring to these papers, Brander and

If the number of domestic firms exceeds the (adjusted) number of foreign firms by more than one, the output of the S home firms in the unconstrained (*laissez-faire*) Cournot equilibrium is larger than the output of each if they were operated by a Stackelberg leader ($q^* > \bar{q}^K$) as is illustrated in Figure 1A. A tax induces a contraction in their outputs and therefore increases their profit (inclusive of the rebated taxes), provided the output of each firm is not reduced below γ_1 . An export subsidy of any magnitude, on the other hand, would adversely affect the home country.

If instead the number of domestic firms exceeds the (adjusted) number of foreign firms by less than one, the output of the S firms in the unconstrained Cournot equilibrium is smaller than if they were operated by a Stackelberg leader ($q^* < \bar{q}^K$), as is illustrated in Figure 1C. A subsidy induces an expansion in their outputs and therefore increases their profits (net of the subsidy payments) provided it does not induce output of each home firm to exceed γ_2 . In this case, of course, the trade policy optimal for the S firms is an export subsidy sufficient to induce each firm to produce \bar{q}^K units.¹⁶ An export tax of any magnitude would, on the other hand, adversely affect the home country.

Suppose finally that the subset now represents firms that are part of a merger or, equivalently, members of a cartel (with perfect enforcement). In general, merging or forming a cartel will cause price to rise, aggregate output to fall, the outputs and profits of the $N - S$ outside firms to rise, and the outputs of the merged entity to fall. Salant et al. (1983) and others have pointed out, however, that the profits of the merged entity (or cartel) may fall. A merger would, of course, increase profits if output of the $N - S$ outside firms remained unchanged.

However, their output will not remain unchanged; it will increase. As a result of outsider expansion, merging or cartelizing some subset of firms exogenously may be disadvantageous; if the merger decision were endogenized, such mergers would presumably not occur.

This "losses from merger" result is yet another consequence of properties of the Cournot model investigated in Section I. The premerger equilibrium is simply the "unconstrained Cournot equilibrium" of that section. As for the postmerger equilibrium, as long as $C'' \geq 0$, each of the S firms party to the merger will produce equal outputs, and the other $N - S$ firms will best reply.

Whenever the number of firms in a merger exceeds the (adjusted) number of outside firms by less than one, the output in the initial unconstrained Cournot equilibrium of each of the S firms that will merge is smaller than the output a Stackelberg leader would choose, as is illustrated in Figure 1C. Since a merger will cause each of the S firms to reduce its output, their profits must fall.

Whenever the number of firms in a merger exceeds the (adjusted) number of outside firms by more than one, a marginal contraction of output is profitable. However, a merger results in a *nonmarginal* contraction in output, and this may or may not be profitable. This case is illustrated in Figure 1A. Let q_m denote the postmerger equilibrium output of each merged firm. If $q_m < \gamma_1$, the merger will cause a loss.¹⁷ If, on the other hand, $\gamma_1 < q_m < q^*$, the merger will cause a gain.

To emphasize the relationship of the seemingly dissimilar literatures on export

Spencer (1985 p. 85 [footnote 6]) acknowledge that "adding more domestic firms weakens the case for domestic subsidies." Indeed, it can easily destroy the case.

¹⁶When the domestic sector consists of a single firm ($S = 1, N \geq 2$), this case must necessarily arise.

¹⁷A particularly striking example has been worked out by Ulrich Zachau (1987). With quadratic costs and proportional demand ($p = v/Q$), for any N and any $1 < S < N$, a merger causes a loss; the only merger that is profitable is merger to monopoly. In this example, whenever a marginal contraction would be profitable, the merger results in a nonmarginal contraction that is so large that $q_m < \gamma_1$, and a loss results.

subsidies and on horizontal mergers, we state the following propositions:

PROPOSITION 1: *Whenever an export subsidy would increase the profits of the home country, merging the entire export sector must cause a loss.*

PROPOSITION 2: *Whenever merging the entire export sector would cause a gain, an export subsidy must reduce the profits of the home country.*

III. Generalizations

In the previous sections, we analyzed the case of quantity competition among firms selling perfect substitutes. We showed that whether an exogenous marginal contraction of the scalar strategies of each of S firms in a subset increases or decreases their profits depends on the number of firms in the constrained subset versus the number of firms in its complement.

This result clearly does not depend on the assumption of quantity competition. For, as Hugo Sonnenschein (1968) pointed out, any result in a market where firms compete in quantity and sell perfect substitutes also holds in a market where firms compete in price and sell perfect complements.¹⁸

Indeed, even in the absence of perfect substitutes (or complements) and even when the strategic variable is neither quantity nor price, analogous results hold. As the Appendix clarifies, such results hold if and only if the interaction involves strategic substitutes.¹⁹

¹⁸One example of the latter form of competition would be independent railroads whose tracks adjoin end to end. Since customers shipping goods from one end to the other would have to pay the *sum* of the fares, this case involves price competition in perfect complements. A theoretical discussion of the profitability of partial consolidation in this example is contained in Gaudet and Salant (1991a); Christopher A. Velturo (1988) and the references he cites discuss the history of railroad consolidations in the United States.

¹⁹Carl Davidson and Raymond Deneckere (1985) implicitly illustrate the "only if" part of the assertion. They consider a model involving price competition and substitutes in demand and show that mergers never

APPENDIX

Generalization of the Comparative-Statics Result on Profits

Consider a symmetric game in which each of N players simultaneously selects a scalar strategy, a_i , $i = 1, \dots, N$, and payoffs are collected. The payoff of player i , $i = 1, \dots, N$, is given by $\pi^i(a_1, \dots, a_i, \dots, a_N)$. We assume that π^i is twice continuously differentiable. We also assume that there exists a unique pure-strategy Nash equilibrium and that it is symmetric. Denote it as $a_i = a$, $i = 1, \dots, N$.

A subset of S players has the constraint $a_i = \bar{a}$ exogenously imposed on it. Without loss of generality, we may assign the subscripts $1, 2, \dots, S$ to these players. The initial equilibrium will be displaced if and only if $\bar{a} \neq a$. We refer to the situation where $\bar{a} = a$ as the unconstrained equilibrium. Given N and S , at the new equilibrium, we will have $a_i(\bar{a})$, $i = S+1, \dots, N$, as an implicit solution to the first-order condition of the $N-S$ players outside the subset.

The effect of a marginal change in \bar{a} on the equilibrium profit of the typical firm in the subset ($i = 1, \dots, S$) is given by

$$(14) \quad \frac{d\pi^i}{d\bar{a}} = \frac{\partial \pi^i}{\partial a_i} + \sum_{\substack{j=1 \\ j \neq i}}^S \frac{\partial \pi^i}{\partial a_j} + \sum_{j=S+1}^N \frac{\partial \pi^i}{\partial a_j} \cdot \frac{da_j}{d\bar{a}}.$$

When $da_j/d\bar{a}$ is positive (negative), we have "strategic complements" (strategic substitutes). We consider the case in which $\partial \pi^i/\partial a_j$ has the same sign at the unconstrained equilibrium for all $j \neq i$, $i = 1, \dots, S$ and $da_j/d\bar{a}$ has the same sign for all $j = S+1, \dots, N$.

cause losses. Strategic complementarity—not the fact that price is the strategic variable—is responsible for this result.

When evaluated at the unconstrained equilibrium, $\partial\pi^i/\partial a_i = 0$ for all $i = 1, \dots, N$. Hence, for strategic complements, (14) implies $\text{sgn}(d\pi^i/d\bar{a}) = \text{sgn}(\partial\pi^i/\partial a_j)$ in a neighborhood of the unconstrained equilibrium, independently of N and S . On the other hand, for strategic substitutes, $\text{sgn}(d\pi^i/d\bar{a})$ clearly depends on (N, S) . For example, when $S = 1$, the first summation is zero and $\text{sgn}(d\pi^i/d\bar{a}) = -\text{sgn}(\partial\pi^i/\partial a_j)$. When $N = S$, the second summation is zero and $\text{sgn}(d\pi^i/d\bar{a}) = \text{sgn}(\partial\pi^i/\partial a_j)$.

These results are general. They include price and quantity competition as important special cases. Hence, $\partial\pi^i/\partial a_j < 0$ arises in quantity competition if the goods are substitutes in demand, and in price competition if the goods are complements in demand; $\partial\pi^i/\partial a_j > 0$ arises in quantity competition if the goods are complements, and in price competition if the goods are substitutes.

REFERENCES

- Brander, James A. and Spencer, Barbara J., "Export Subsidies and International Market Share Rivalry," *Journal of International Economics*, February 1985, 18, 83-100.
- Bulow, Jeremy I., Geanakoplos, John D. and Klemperer, Paul D., "Multimarket Oligopoly: Strategic Substitutes and Complements," *Journal of Political Economy*, June 1985, 93, 488-511.
- Carter, Colin A., Hueth, Darrell L., Mamer, John W. and Schmitz, Andrew, "Labor Strikes and the Price of Lettuce," *Western Journal of Agricultural Economics*, July 1981, 5, 1-14.
- Deneckere, Raymond and Davidson, Carl, "Incentives to Form Coalitions with Bertrand Competition," *Rand Journal of Economics*, Winter 1985, 16, 473-86.
- Dixit, Avinash K., "International Trade Policy for Oligopolistic Industries," *Economic Journal*, December 1984, 94 (Supplement), 1-16.
- Eaton, Jonathan and Grossman, Gene M., "Optimal Trade and Industrial Policy under Oligopoly," *Quarterly Journal of Economics*, May 1986, 101, 383-406.
- Farrell, Joseph and Shapiro, Carl, "Horizontal Mergers: An Equilibrium Analysis," *American Economic Review*, March 1990, 80, 107-26.
- Gaudet, Gérard, and Salant, Stephen W., "Profitability of Exogenous Output Contractions: A Comparative-Static Analysis with Application to Strikes, Mergers and Export Subsidies," CREST Working Paper 89-09, Department of Economics, University of Michigan, February 1989.
- _____ and _____, (1991a) "Towards a Theory of Horizontal Mergers," in George Norman and Manfredi La Manna, eds., *The New Industrial Economics: Recent Developments in Industrial Organization, Oligopoly and Game Theory*, Cheltenham, U.K.: Edward Elgar Publishing, 1991 (forthcoming).
- _____ and _____, (1991b) "Uniqueness of Cournot Equilibrium: New Results from Old Methods," *Review of Economic Studies*, 1991 (forthcoming).
- Salant, Stephen W., "Export Subsidies as Instruments of Economic and Foreign Policy," Rand Note N.2120-USDP, Santa Monica, CA: The Rand Corporation, 1984.
- _____, Switzer, Sheldon and Reynolds, Robert J., "Losses from Horizontal Merger: The Effects of an Exogenous Change in Industry Structure on Cournot-Nash Equilibrium," *Quarterly Journal of Economics*, February 1983, 98, 185-99.
- Sonnenschein, Hugo, "The Dual of Duopoly is Complementary Monopoly: Or, Two of Cournot's Theories are One," *Journal of Political Economy*, March/April 1968, 76, 316-8.
- Velluro, Christopher A., "Achieving Cost Efficiencies Through Merger: Evidence from the U.S. Rail Industry," mimeo, Massachusetts Institute of Technology, 1988.
- Zachau, Ulrich, "Mergers in the Model of an R&D-Race," mimeo, Universitaet Bonn, Institut fuer Gesellschafts-und Wirtschaftswissenschaften, 1987.

Pricing Schemes and Cournotian Equilibria

By C. D'ASPREMONT, R. DOS SANTOS FERREIRA, AND
L.-A. GÉRARD-VARET*

There are, essentially, two basic models for studying oligopolistic competition: the Cournot model, with quantity-setting firms in a market for a single homogeneous good, and the monopolistic competition model, with price-setting firms, each in a different market of a differentiated good. In both models, the analysis starts by assuming a large set of consumers who adjust in a perfectly competitive way. Only the firms are supposed to behave strategically and to privilege one strategic variable, the quantity produced or the selling price. This paper is an attempt—among others¹—to integrate the two models and to admit the interaction of the two kinds of strategic variables.

This integration will be based on a reinterpretation of Augustin Cournot's approach. In the famous chapter in which Cournot (1838) introduced his oligopoly theory, he does mention, from the start, that the selling price should "necessarily" be the same for each producer. However, he does not elaborate much on price formation, except to say that it is convenient to use the inverse demand function. One may argue that the lack of explanation for price determination in the Cournot model is analogous to the lack of one in the perfect-competition model, and a Walrasian auctioneer might be hypothetically introduced. Of

course, one difference is that the auctioneer may only act on one side of the market, the demand side. Supply is dictated to him by the firms, and so they themselves integrate into their computations the influence of their supply on the resulting market price. However, the lack of a price-adjustment theory and the idea that such a theory should be rooted on a price-adjustment process under monopoly (one firm adjusting both price and quantity), as advocated by Kenneth J. Arrow (1959) for the perfect-competition model, apply as well to the Cournot oligopoly model. To invoke the "necessity" that the producers of a homogeneous good charge the same price is not different from invoking the "impersonal forces of the market" to justify one competitive equilibrium price that all agents take as given. The argument that, out of equilibrium, producers behave as monopolists and do consider the influence of the price they charge on the (imperfectly elastic) demand they face should also be maintained. The main contrast, finally, is that this argument is still valid at equilibrium in the Cournot oligopoly model. Even there, the producers do not take the price they charge as given. Moreover, they know that, if they were charging different prices, some adjustment process would take place leading again to a single market price. Out of equilibrium, there is a discrepancy between the price they want to charge and the market price (or the average price²), and this triggers the adjustment. At equilibrium the two coincide.

In this paper, we shall not introduce a theory of price adjustment under conditions

*CORE, Université Catholique de Louvain, Louvain-la-Neuve, Belgium; BETA, Université Louis Pasteur, Strasbourg, France; and GREQE, École des Hautes Études en Sciences Sociales, Marseille, France. We thank Jean Gabszewicz, Ehud Kalai, Philippe Michel, Louis Philips, Ya'ir Tauman, and Xavier Vives for their comments and suggestions. Support from the Commissariat Général du Plan and the Ministère de la Recherche is also gratefully acknowledged.

¹See, for example, Sanford J. Grossman (1981), David Kreps and José A. Scheinkman (1983), Jean Frayssé (1986), Xavier Vives (1986), James W. Friedman (1987), and Paul Klemperer and Margaret Meyer (1989).

²To quote Arrow (1959 p. 48), "However the 'price' whose movements are explained by the law must be thought of as the average price."

of imperfect competition. We shall insist instead that the result of any such theory would be different from the result of a theory of price adjustment under perfect competition: at equilibrium each firm should be fully aware of its own influence on the market price. To formalize this conclusion, we shall introduce explicitly a concept of "pricing scheme" associating with a vector of price announcements the resulting market price. It will appear that, if the pricing scheme (which is nothing else than a coordination device) is sufficiently responsive to individual price signals, then we get the Cournot equilibrium. This leads to the interpretation of a Cournot equilibrium as the coordinated optimal decisions of a set of monopolists, each facing some (imperfectly elastic) residual demand. In the original Cournot model, the same coordination is ensured by the use of the inverse demand function. Formally, pricing schemes have the same status as auctions or bidding mechanisms. They could be assimilated to what is known in industrial organization as "facilitating practices" (see Steven C. Salop, 1985): these are more or less explicit customs established in some industries to allow for price coordination. Examples are the best-price guarantee given to a customer, either with respect to other sellers ("meet-or-release" clause) or with respect to other customers ("most-favored-customer" clause) and the practice of public advance notification of price increases.³ However, we do not introduce pricing schemes here as corresponding to well-specified price-formation mechanisms used in particular industries. They are seen as an explicit but formal representation of the coordination of pricing

decisions that is reached in an industry for a homogeneous product (maybe after a long process) and which is implicit in the use of the inverse demand function in Cournot's traditional approach.

Furthermore, by introducing pricing schemes, we allow for a more natural definition of Cournotian oligopolistic competition with several sectors, each containing several producers of the same homogeneous good, the market price in any one sector being fixed through its own respective pricing scheme. We thus get, in each sector, a well-defined juxtaposition of monopoly problems, one for each firm in the sector, contingent on the total quantity produced in the same sector and on the prices set in the other sectors. The benefit is to avoid assuming that the producers are able to carry over in their computation the inversion of a complete demand system, taking all cross-sectoral effects into account. In a general equilibrium model, this would lead to an alternative to the Cournot-Walras approach (as developed by Jean J. Gabszewicz and Jean-Philippe Vial [1972]) and to the market game approach (see e.g., Lloyd S. Shapley and Martin Shubik, 1977; Pradeep Dubey, 1981; Leo K. Simon, 1984).

The paper is organized as follows. In Section I, a formal concept of equilibrium with pricing schemes is compared to the Cournot equilibrium and then, in Section II, it is illustrated for a particular scheme, the min-pricing scheme. In Section III, it is extended to the multisectoral case generalizing both the Cournot and the monopolistic competition concepts.

I. Pricing Schemes in the Cournot Model

In this section, we consider the market for a single homogeneous good with an extended real-valued demand function D , defined on \mathbb{R}_+ , and a set $N = \{1, \dots, i, \dots, n\}$ of firms. Each firm $i \in N$ can produce any quantity $y_i \geq 0$, at a nonnegative cost $C_i(y_i)$, and choose any price signal $\psi_i \geq 0$. Each C_i is a continuous increasing function on \mathbb{R}_+ . As discussed in the introduction, the market price is supposed to be determined by a pricing scheme P , a continuous nondecreas-

³The main result of the facilitating-practice literature is to show how such clauses can implement prices above the competitive price. For example Ehud Kalai and Mark A. Satterthwaite (1986) and Christopher Doyle (1988) get the implementation of the collusive price. However, by introducing discount possibilities below list prices in a second stage, Charles A. Holt and David T. Scheffman (1987) get the Cournot price as the maximal implementable price. For an experimental approach, see David M. Grether and Charles R. Plott (1984).

ing function from \mathbb{R}_+^n to \mathbb{R}_+ , associating with each vector of price signals $\psi = (\psi_1, \dots, \psi_i, \dots, \psi_n)$ a single price $P(\psi)$. For a given pricing scheme P , we thus obtain a game involving the n firms, the strategies of firm i being the set of nonnegative quantity-price pairs (y_i, ψ_i) and, for any vector (\mathbf{y}, ψ) of such strategies, the payoff of firm i being given by the profit function

$$\Pi_i(\mathbf{y}, \psi) \equiv y_i P(\psi) - C_i(y_i).$$

In addition, a feasibility constraint is imposed on (\mathbf{y}, ψ) in the strategy set:

$$Y \equiv \sum_{i=1}^n y_i \leq D(P(\psi)).$$

Then, letting $\mathbf{y}_{-i} \equiv (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n) \in \mathbb{R}_+^{n-1}$ (and defining ψ_{-i} similarly) and also $Y_{-i} \equiv \sum_{j \neq i} y_j$, we define a P -equilibrium as a vector (\mathbf{y}^*, ψ^*) in \mathbb{R}_+^{2n} , such that $Y^* = D(P(\psi^*))$ and, for every $i \in N$, (y_i^*, ψ_i^*) is a solution to

$$\max_{(y_i, \psi_i)} y_i P(\psi_i, \psi_{-i}^*) - C_i(y_i)$$

such that $y_i \leq D(P(\psi_i, \psi_{-i}^*)) - Y_{-i}^*$, $y_i \geq 0$, $\psi_i \geq 0$. To find the Cournot equilibrium in this way, it is clear that pricing schemes have to be more precisely specified. Three classes will be discussed, varying according to the degree of control each firm has on the market price. They are schemes such that (1) all firms together exercise complete control, (2) each individual firm has local control, or (3) each individual firm has complete control on the market price. Property (1) is minimal and amounts to requiring that P has full range: $P(\mathbb{R}_+^n) = \mathbb{R}_+$. However, the main feature distinguishing Cournotian competition from perfect competition is the influence an individual firm may have on the market price. Hence, property (2) requires that the pricing scheme P be strictly increasing in each variable ψ_i . Abstract pricing schemes satisfying these first two properties are, for example, the "arithmetic mean" $\bar{P}(\psi) = (\sum_{i=1}^n \psi_i)/n$, and the "harmonic mean" $\hat{P}(\psi) = n / [\sum_{i=1}^n (1/\psi_i)]$ if $\psi \gg 0$ ($\hat{P}(\psi) = 0$, otherwise). By contrast, the "min-pricing scheme" $P^{\min}(\psi) = \min_j \{\psi_j\}$ and the "max-

pricing scheme" $P^{\max}(\psi) = \max_j \{\psi_j\}$, which have full range, are not strictly increasing in ψ_i . One may go further and consider as property (3) that, for every $i \in N$ and $\psi_{-i} \in \mathbb{R}_+^{n-1}$, $P(0, \psi_{-i}) = 0$ and, for $\psi_i \gg 0$, $P(\cdot, \psi_{-i})$ has full range. This property is not satisfied by the previous examples: the arithmetic mean and the max-pricing scheme give complete control to the individual firms only "upwards"; the harmonic mean and the min-pricing scheme give complete control only "downwards." A pricing scheme satisfying all three properties is the "geometric mean" $\bar{P}(\psi) = [\prod_{i=1}^n \psi_i]^{1/n}$. We shall denote by \mathcal{P}_1 , \mathcal{P}_2 , and \mathcal{P}_3 the sets of pricing schemes satisfying respectively the first, the first two, and all three properties. The interest in considering these classes of pricing schemes is that they imply a close relationship between P -equilibria and Cournot equilibria. This is exhibited in Proposition 1 below.

Recall that, whenever the inverse demand function D^{-1} is well defined [i.e., continuous, decreasing on \mathbb{R}_+ , and such that, for $0 < Y < D(0)$, $D^{-1}(Y) = p$ if and only if $D(p) = Y$] and nontrivial [i.e., $D^{-1}(Y) > 0$ for some $Y > 0$] a Cournot equilibrium is a quantity vector $\mathbf{y}^c \in \mathbb{R}_+^n$ such that, for every $i \in N$, y_i^c is a solution to

$$\max_{y_i \geq 0} y_i D^{-1}(Y_{-i}^c + y_i) - C_i(y_i)$$

$$\text{with } Y_{-i}^c = \sum_{j \neq i} y_j^c.$$

We may then show the following.

PROPOSITION 1: *Let the inverse demand function D^{-1} be well defined and nontrivial. We have:*

- (a) *For any pricing scheme $P \in \mathcal{P}_1$, if \mathbf{y}^c is a Cournot equilibrium, then (\mathbf{y}^c, ψ^c) is a P -equilibrium where ψ^c is chosen so that $P(\psi^c) = D^{-1}(Y^c)$.*
- (b) *For any pricing scheme $P \in \mathcal{P}_3$, if (\mathbf{y}^*, ψ^*) is a P -equilibrium, then \mathbf{y}^* is a Cournot equilibrium.*
- (c) *Assume that the profit $[y_i D^{-1}(Y_{-i} + y_i) - C_i(y_i)]$ is a strictly quasi-concave function of y_i for every $Y_{-i} \in \mathbb{R}_+$ and every $i \in N$. For any pricing scheme $P \in \mathcal{P}_2$, if*

(\mathbf{y}^*, Ψ^*) is a P -equilibrium, then \mathbf{y}^* is a Cournot equilibrium.

PROOF:

(a) Consider a Cournot equilibrium \mathbf{y}^c and take any $P \in \mathcal{P}_1$. Because P has full range, we can pick $\Psi^c \in \mathbb{R}_+^n$ such that $P(\Psi^c) = D^{-1}(Y^c)$. Suppose (\mathbf{y}^c, Ψ^c) is not a P -equilibrium; that is, $y_i^c P(\psi_i^0, \Psi_{-i}^c) - C_i(y_i^c) > y_i^c P(\Psi^c) - C_i(y_i^c)$ and $y_i^0 \leq D(P(\psi_i^0, \Psi_{-i}^c)) - Y_{-i}^c$, for some $i \in N$, $\psi_i^0 \geq 0$ and $y_i^0 \geq 0$. Since D^{-1} is well defined, the second inequality implies

$$D^{-1}(Y_{-i}^c + y_i^0) \geq P(\psi_i^0, \Psi_{-i}^c)$$

so that,

$$\begin{aligned} y_i^0 D^{-1}(Y_{-i}^c + y_i^0) - C_i(y_i^0) \\ > y_i^c D^{-1}(Y^c) - C_i(y_i^c) \end{aligned}$$

which is a contradiction to \mathbf{y}^c being a Cournot equilibrium.

(b) For $P \in \mathcal{P}_3$, suppose (\mathbf{y}^*, Ψ^*) is a P -equilibrium but \mathbf{y}^* is not a Cournot equilibrium; that is, for some $y_i^0 \geq 0$ and $i \in N$,

$$\begin{aligned} y_i^0 D^{-1}(Y_{-i}^* + y_i^0) - C_i(y_i^0) \\ > y_i^* D^{-1}(Y^*) - C_i(y_i^*) \end{aligned}$$

and

$$D^{-1}(Y^*) = P(\Psi^*) > 0.$$

Since i has complete control, there is $\psi_i^0 \geq 0$ such that $P(\psi_i^0, \Psi_{-i}^*) = D^{-1}(Y_{-i}^* + y_i^0)$ and

$$\begin{aligned} y_i^0 P(\psi_i^0, \Psi_{-i}^*) - C_i(y_i^0) \\ > y_i^* P(\Psi^*) - C_i(y_i^*) \end{aligned}$$

which contradicts that (\mathbf{y}^*, Ψ^*) is a P -equilibrium.

(c) If $P \in \mathcal{P}_2$, the problem with the previous argument is that one cannot be sure to find $\psi_i^0 \geq 0$ such that $P(\psi_i^0, \Psi_{-i}^*) = D^{-1}(Y_{-i}^* + y_i^0)$. However, by strict quasi-concavity, one can find some y_i^0 satisfying the above strict inequality arbitrarily close

to y_i^* and, hence, some ψ_i^0 such that $P(\psi_i^0, \Psi_{-i}^*) = D^{-1}(Y_{-i}^* + y_i^0)$, since $P(\cdot, \Psi_{-i}^*)$ is strictly increasing. The result follows.

Now, given a pricing scheme P and a P -equilibrium (\mathbf{y}^*, Ψ^*) , we denote by B_i^* the set of prices

$$\arg \sup_{p \geq 0} \{p[D(p) - Y_{-i}^*] - C_i(D(p) - Y_{-i}^*)\}$$

subject to

$$D(p) - Y_{-i}^* \geq 0.$$

This set may include $p = \infty$. We see that B_i^* is the set of prices among which firm i would choose if it were a monopolist facing the residual demand $[D(p) - Y_{-i}^*]$. We shall call firm i a " P -leader" at (\mathbf{y}^*, Ψ^*) whenever $P(\Psi^*) \in B_i^*$. A firm i that is not a P -leader at (\mathbf{y}^*, Ψ^*) will be called a " P -follower." Then, y_i^* is a solution to

$$\max_{y_i \geq 0} P(\Psi^*) y_i - C_i(y_i)$$

subject to

$$y_i \leq D(P(\Psi^*)) - Y_{-i}^*.$$

We see that a P -follower⁴ can only be a price-taker with respect to the price $P(\Psi^*)$. The next proposition shows how the Cournot equilibrium can be interpreted. At a Cournot equilibrium, each firm behaves as a monopolist facing the residual demand.

PROPOSITION 2: *If the inverse demand function D^{-1} is well defined and nontrivial, then, for any full-range pricing scheme ($P \in \mathcal{P}_1$) and any P -equilibrium (\mathbf{y}^*, Ψ^*) , \mathbf{y}^* is a Cournot equilibrium if and only if all firms are P -leaders at (\mathbf{y}^*, Ψ^*) .*

PROOF:

Suppose first that \mathbf{y}^* is not a Cournot equilibrium. Then, for some $i \in N$ and

⁴ If i is a P -leader, then y_i^* is also a solution to this program; otherwise (\mathbf{y}^*, Ψ^*) would not be a P -equilibrium. The difference is that the leader chooses an optimal price.

$$y_i^0 \geq 0,$$

$$\begin{aligned} & y_i^0 D^{-1}(Y_{-i}^* + y_i^0) - C_i(y_i^0) \\ & > y_i^* D^{-1}(Y^*) - C_i(y_i^*) \\ & \text{with } D^{-1}(Y^*) = P(\Psi^*). \end{aligned}$$

Then, taking $p^0 = D^{-1}(Y_{-i}^* + y_i^0)$, we see immediately that i cannot be a P -leader. Now, if y^* is a Cournot equilibrium, then (y^*, Ψ^*) is a P -equilibrium (in the first part of Proposition 1 we need only that P be of full range). Thus, if i is not a P -leader, for some $p^0 \geq 0$, we must have,

$$\begin{aligned} & p^0 [D(p^0) - Y_{-i}^*] - C_i(D(p^0) - Y_{-i}^*) \\ & > P(\Psi^*) [D(P(\Psi^*)) - Y_{-i}^*] \\ & \quad - C_i(D(P(\Psi^*)) - Y_{-i}^*) \end{aligned}$$

or, letting $y_i^0 = D(p^0) - Y_{-i}^* \geq 0$ [and $p^0 = D^{-1}(Y_{-i}^* + y_i^0)$], we obtain the same strict inequality as above [since $y_i^* = D(P(\Psi^*)) - Y_{-i}^*$], contradicting that y^* is a Cournot equilibrium.

This leads to a typology of P -equilibria. Besides the P -equilibrium in which all firms are P -leaders and which is of a Cournot type, we have P -equilibria in which all firms are P -followers and which include the Bertrand-type equilibrium, and we also have those in which some firms are P -leaders and some are P -followers. This is clarified in the following example, in which all types of P -equilibria can be exhibited for a particular pricing scheme, the min-pricing scheme P^{\min} .

II. The Special Case of the Min-Pricing Scheme

Consider two firms, with a linear demand function,

$$D(p) = \max\{0, a - bp\} \quad a > 0, b > 0$$

and for each firm i , the same twice continuously differentiable cost function C_i with $C' > 0$, $C'' > 0$, and $C(0) = C'(0) = 0$. Let

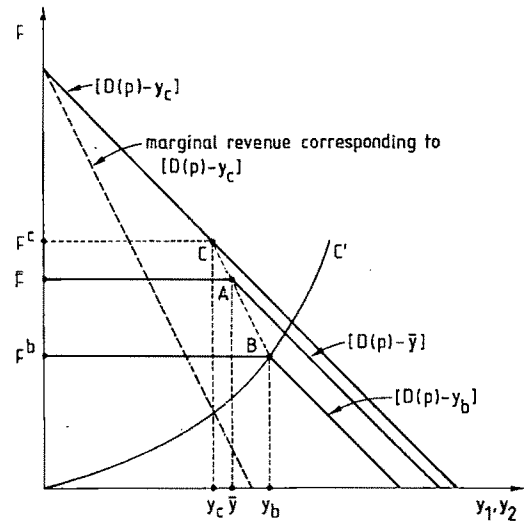


FIGURE 1. SYMMETRIC P^{\min} -EQUILIBRIA

$\eta(p) \equiv -p[D'(p)/D(p)]$ denote the price elasticity. Here $\eta(p) = bp/(a - bp)$. The first-order conditions for a Cournot equilibrium y^c are well known to be

$$p^c \left[1 - \frac{y_i^c}{y_i^c + y_j^c} \frac{1}{\eta(p^c)} \right] = C'(y_i^c)$$

with $y_i^c + y_j^c = D(p^c)$, $i \neq j$. Here, this clearly implies $y_i^c = y_j^c = y_c$ and is simply given by the solution of

$$\frac{1}{b}(a - 3y_c) = C'(y_c).$$

This can be interpreted as a symmetric P^{\min} -equilibrium in which both firms are P -leaders (by Proposition 2), by letting $\psi_1^* = \psi_2^* = p^c$ and $y_1^* = y_2^* = y_c$. This symmetric P^{\min} -equilibrium appears as point C in Figure 1.

There are other symmetric P^{\min} -equilibria for which, on the contrary, all firms are P -followers: each would prefer to increase the price but cannot individually do so. Therefore, each produces as much as possible considering the residual demand, as long as the market price is higher than the marginal cost. It is as if firms were facing "kinked" demand curves, with the

equilibrium corresponding to the “kink,” the point at which there is a discontinuity in the marginal revenue. Two such equilibria are illustrated by points A and B in Figure 1. Point A corresponds to the symmetric P^{\min} -equilibrium with $\psi_1^* = \psi_2^* = \bar{p}$ and $y_1^* = y_2^* = \bar{y}$ and satisfies the conditions

$$\bar{p} \left[1 - \frac{1}{2} \left(\frac{1}{\eta(\bar{p})} \right) \right] < C'(\bar{y}) < \bar{p}.$$

Point B corresponds to the symmetric P^{\min} -equilibrium with $\psi_1^* = \psi_2^* = p^b$ and $y_1^* = y_2^* = y_b$ and satisfies the conditions

$$p^b \left[1 - \frac{1}{2} \left(\frac{1}{\eta(p^b)} \right) \right] < C'(y_b) = p^b.$$

It is in fact a Bertrand equilibrium. There is a continuum of symmetric P^{\min} -equilibria of the type given by point A between points B and C, corresponding to prices between p^b and p^c . No symmetric P^{\min} -equilibrium corresponds to prices below p^b , since then the equilibrium condition $D(\min\{\psi_1^*, \psi_2^*\}) = Y^*$ cannot be satisfied. Also there are no P^{\min} -equilibria corresponding to prices above the Cournot price p^c .⁵

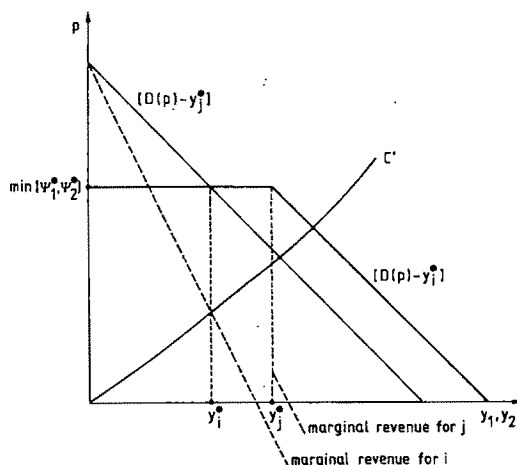
There are also nonsymmetric P^{\min} -equilibria. These would consist of vectors $(\mathbf{y}^*, \boldsymbol{\psi}^*)$ such that, for $p^* = \min\{\psi_1^*, \psi_2^*\}$,

$$p^* \left[1 - \frac{y_i^*}{Y^*} \left(\frac{1}{\eta(p^*)} \right) \right] = C'(y_i^*) \quad \text{for some } i \in \{1, 2\}$$

$$p^* \geq C'(y_j^*) > p^* \left[1 - \frac{y_j^*}{Y^*} \left(\frac{1}{\eta(p^*)} \right) \right]$$

for $j \neq i$.

This implies that $y_j^* > y_i^*$ and is illustrated in Figure 2. This is a “price leadership” type of equilibrium: i is the price-maker, j is the price-taker (as such, firm j could even

FIGURE 2. NONSYMMETRIC P^{\min} -EQUILIBRIA

be at a point where $\min\{\psi_1^*, \psi_2^*\} = C'(y_j^*)$,
but no further).

III. A Simple Multisectoral Extension

The advantage of the approach of oligopoly via pricing schemes is to allow straightforward extension of a Cournot-type equilibrium concept to the multisector case. Indeed, the usual definition of the Cournot equilibrium requires the definition of the inverse demand function to reduce the strategy space of each firm to the choice of a quantity. However, in the multisector case, the demand in each sector depends on the price in the other sectors, and so one has to invert a complete demand system as in the Cournot-Walras approach (see Gabszewicz and Vial, 1972). This implies that even stronger conditions have to be imposed before defining the equilibrium concept. Moreover, the inverse demand system thus obtained might not facilitate the interpretation of the additional restrictions needed later to prove existence. Using pricing schemes, we can define directly the equilibrium concept and postpone the introduction of specific assumptions to the time when the existence problem would have to be examined.

Suppose we have m goods, each one homogeneous, and that the market for each

⁵In their two-stage procedure Holt and Scheffman (1987) reach the same conclusions for the symmetric case (see their proposition 2).

good k is characterized by an extended real-valued demand function D_k , defined for any vector of nonnegative prices $\mathbf{p} \equiv (p^1, \dots, p^k, \dots, p^m) \in \mathbb{R}_+^m$. Also, in each sector k , we have a set N^k of n^k firms. Each firm $i \in N^k$ can produce a quantity $y_i \in \mathbb{R}_+$ of good k and has a continuous increasing cost function $C_i(y_i)$. In the market for good k , the price is determined by a pricing scheme P^k , a continuous nondecreasing function of the price signals of all firms in sector k ; say $\Psi^k = (\psi_i^k)_{i \in N^k} \in \mathbb{R}_+^{n^k}$. Letting $Y^k \equiv \sum_{i \in N^k} y_i$, $Y_{-i}^k \equiv \sum_{j \in N^k \setminus \{i\}} y_j$, for $k = 1, \dots, m$, and $\mathbf{P} = (P^1, \dots, P^k, \dots, P^m)$, we define a P -equilibrium as a vector (\mathbf{y}^*, Ψ^*) in $\mathbb{R}_+^n \times \mathbb{R}_+^n$ such that, for $k = 1, \dots, m$, $Y^k = D^k(P(\Psi^*))$ with $P(\Psi^*) = (P^1(\Psi^*), \dots, P^m(\Psi^*))$, and for every $i \in N^k$, (y_i^*, ψ_i^*) is a solution to

$$\max_{(y_i, \psi_i)} y_i P^k(\psi_i, \Psi_{-i}^*) - C_i(y_i)$$

subject to

$$y_i \leq D^k(P(\psi_i, \Psi_{-i}^*)) - Y_{-i}^k, y_i \geq 0, \psi_i \geq 0.$$

From the preceding section, we know that to get Cournot-type equilibria we have to require at least that $P_k \in \mathcal{D}_2$ for every k . In fact, the equilibrium we obtain then can be described without referring to any pricing scheme. Indeed, it amounts to having a pair of prices and quantities $(\mathbf{p}^*, \mathbf{y}^*)$ in $\mathbb{R}_+^m \times \mathbb{R}_+^n$ such that each firm i in each sector k is (at least locally) a P^k -leader: it behaves as a price-setting monopolist facing the demand function $[D^k(\cdot, \mathbf{p}^{-k*}) - Y_{-i}^{k*}]$ which is contingent on the other equilibrium prices \mathbf{p}^{-k*} and the equilibrium total quantity of good k , Y_{-i}^{k*} , produced by the other firms of sector k . This equilibrium concept can be defined independently from pricing schemes and may be called a *Cournotian monopolistic competition equilibrium*. Thus, for this equilibrium concept, pricing schemes appear simply as formal devices allowing one to model the coordination between producers who are all conscious of their own influence on their market price, without computing the inverse of the complete demand system.

IV. Conclusion

In this paper, we have proposed a reinterpretation of Cournot equilibrium based on a general and formal coordination device, the pricing scheme, satisfying some global and individual responsiveness properties. The exact determination of this device is not required. For the multisectoral case, we get an equilibrium concept that generalizes both the Cournot equilibrium and the monopolistic competition equilibrium (i.e., if there is one sector, it generalizes Cournot; if there is only one firm per sector, it generalizes monopolistic competition). The more general equilibrium concept to which it leads, the Cournotian monopolistic competition equilibrium, can be described as the solution to the juxtaposition of many monopoly problems, each monopolist facing a demand function contingent on the equilibrium quantities produced in its sector and on the equilibrium prices set in the other sectors. We leave for future work the existence problem for such a concept in a general-equilibrium framework.

REFERENCES

- Arrow, Kenneth J., "Toward a Theory of Price Adjustment," in M. Abramovitz and others, eds., *The Allocation of Economic Resources*, Stanford: Stanford University Press, 1959.
- Cournot, Augustin, *Recherches sur les Principes Mathématiques de la Théorie des Richesses*, Paris: Hachette, 1838; English translation (N. Bacon, translator) *Researches into the Mathematical Principles of the Theory of Wealth*. New York: Macmillan, 1897.
- Doyle, Christopher, "Different Selling Strategies in Bertrand Oligopoly," *Economics Letters*, 1988, 28 (4), 387-90.
- Dubey, Pradeep, "Price-Quantity Strategic Market Games," *Econometrica*, January 1981, 50, 111-26.
- Frayssé, Jean, *Equilibres de Cournot dans les Grands Marchés*, Monographies d'Econometrie, Paris: Editions du CNRS, 1986.
- Friedman, James W., "On the Strategic Importance of Prices versus Quantities,"

- Discussion Paper, University of North Carolina, 1987.
- Gabszewicz, Jean J. and Vial, Jean-Philippe, "Oligopoly 'à la Cournot' in General Equilibrium Analysis," *Journal of Economic Theory*, June 1972, 4, 381-400.
- Grether, David M. and Plott, Charles R., "The Effects of Market Practices in Oligopolistic Markets: An Experimental Examination of the Ethyl Case," *Economic Inquiry*, October 1984, 22, 479-507.
- Grossman, Sanford J., "Nash Equilibrium and the Industrial Organizational Markets with Large Fixed Costs," *Econometrica*, September 1981, 49, 1149-72.
- Holt, Charles A. and Scheffman, David T., "Facilitating Practices: The Effects of Advance Notice and Best-Price Policies," *Rand Journal of Economics*, Summer 1987, 18, 187-97.
- Kalai, Ehud and Satterthwaite Mark A., "The Kinked Demand Curve, Facilitating Practices, and Oligopolistic Competition," MEDS Discussion Paper No. 677, Northwestern University, February 1986.
- Klemperer, Paul and Meyer, Margaret, "Supply Function Equilibria in Oligopoly under Uncertainty," *Econometrica*, November 1989, 99, 1243-77.
- Kreps, David and Scheinkman, José A., "Quantity Precommitment and Bertrand Competition Yield Cournot Outcomes," *Bell Journal of Economics*, Autumn 1983, 14, 326-37.
- Salop, Steven C., "Practices that (Credibly) Facilitate Oligopoly Coordination," in Joseph E. Stiglitz and G. Frank Mathewson, eds., *New Developments in Market Structure*, Cambridge: Cambridge University Press, 1985.
- Shapley, Lloyd S. and Shubik, Martin, "Trade Using One Commodity as a Means of Payment," *Journal of Political Economy*, October 1977, 85, 937-68.
- Simon, Leo K., "Bertrand, the Cournot Paradigm and the Theory of Perfect Competition," *Review of Economic Studies*, April 1984, 51, 209-30.
- Vives, Xavier, "Commitment, Flexibility and Market Outcomes," *International Journal of Industrial Organization*, June 1986, 4, 217-29.

Redistribution and Capital Formation

By PETER RANGAZAS*

The conventional view regarding the macroeconomic effects of redistribution from the rich to the poor is that it will lower aggregate savings and physical capital accumulation. Savings will fall because the marginal propensity to save is thought to be higher for the rich than for the poor. Working from a life-cycle framework, Laurence Kotlikoff (1984 p. 1598) pointed out that there is little empirical support for this view and, in fact, there are several *a priori* reasons to believe it may *not* be true. Alan Blinder (1975) extended the life-cycle model to include a taste-for-bequest motive and examined the same issue. He showed that, if households systematically differ solely according to their permanent incomes, then aggregate savings will fall with lump-sum redistribution only if bequests are a luxury good; otherwise, savings may remain the same or rise.

One argument in the conventional view's favor is based on models of altruistically motivated intergenerational transfers.¹ In these models, rich families, who can afford to make efficient levels of human capital investments in their children, typically have enough wealth to make sizable bequests of financial assets. Poorer families, on the other hand, make their planned bequests exclusively in the form of human capital expenditures. No planned financial bequests are

made by these families because the return on the inefficiently low levels of human capital exceeds that on financial assets. In this case, a marginal redistribution of wealth would apparently lower aggregate savings, due to a reduction in financial bequests made by the rich with no compensating increase by the poor. Human capital production, however, would apparently increase. Transfers from rich to poor families would not alter the efficient investment in rich children, since for these families only financial bequests would be affected at the margin, but such transfers would raise human capital investment in poorer families (Gary Becker, 1981 pp. 121–2). While this argument in support of the conventional view has recently been emphasized by several authors (Blinder, 1976; Becker, 1981; Becker and Nigel Tomses, 1986), Paul Menchik and Martin David (1983) have traced it all the way back to Marshall's *Principles of Economics*.

In this paper, I show that the argument in favor of the conventional wisdom is only valid for a temporary (one generation) policy of redistribution, with fixed factor prices. Altering either the temporary nature of the policy or the fixed-price assumption allows the conclusions to be reversed under fairly general conditions. Relaxing the fixed-price assumption is a natural extension of the existing analysis which needs little defense. The distinction between permanent and temporary policies also clearly becomes important if families remain unconstrained (or constrained) for more than one generation.²

*Department of Economics, Indiana University-Purdue University at Indianapolis, Indianapolis, IN 46202. I am grateful to William Lord for useful discussions on the topic and to Sharon Zehr Rangazas for her assistance and comments. The paper also greatly benefited from the many thoughtful suggestions offered by the referees.

¹The extent to which any model focusing on bequests is empirically relevant for capital formation depends on the role of bequests motives relative to life-cycle factors in determining aggregate savings (see the recent debate on this topic between Franco Modigliani [1988] and Kotlikoff [1988]). For criticisms of the importance of altruistic bequests in particular, see Bernheim (1989 pp. 66–7).

²This will depend, in part, on how fast families regress toward the mean of wealth. Jere Behrman and Paul Taubman (1985) and Becker and Tomses (1986) find strong regression toward the mean in earnings. On the other hand, Menchik (1979) finds a great deal of wealth immobility for rich households, and Gary Solon et al. (1987) produce evidence suggesting substantial immobility for low-income families. Christopher Ruhm (1988) presents a theoretical model in which a family

For example, if an unconstrained household has a unit of wealth taxed away but also knows the policy applies to its children, then bequests savings will not necessarily fall. The outcome depends on the size of opposing marginal propensities to bequeath out of parents' and children's lifetime wealth.

I. The Microeconomic Model

Consider a household, headed by parents who live for one period.³ In each period, the parents choose lifetime consumption (c_t), expenditures per child (x_t), and a financial bequest (B_t), which is divided evenly among the children.⁴ The financial bequest is the amount by which parents plan to augment the wealth of their children, net of taxes and interest payments. The portion of lifetime wealth that must be set aside to accomplish this transfer is defined to be $p_t B_t$, where $p_t = (1 + r_{t+1})^{-1}$ may be interpreted as the price of making a transfer of net wealth to children.

There are $n_{t+1} = (1 + g)$ identical children per household, where g is the population growth rate. The household receives satisfaction from its own consumption and the total lifetime wealth of its children. The problem confronting the household is then

to maximize the following utility function:

$$(1) \quad U = u(c_t, W_{t+1} n_{t+1})$$

subject to its lifetime wealth constraint

$$(2) \quad c_t + n_{t+1} x_t + p_t B_t = W_t$$

where $W_t = I_t + R_t - T_t + w_t h_t$, in which I_t is inherited wealth, R_t is a lump-sum transfer made by the government, T_t is a lump-sum tax, and $w_t h_t$ is lifetime earnings. The variable w_t is the wage rate, and h_t is the stock of human capital. This stock is a function of last period's parental expenditures on education, x_{t-1} :

$$(3) \quad h_t = h(x_{t-1}).$$

The first-order conditions are

$$(4a) \quad u_1 = \lambda_t$$

$$(4b) \quad u_2 w_{t+1} h' = \lambda_t$$

$$(4c) \quad u_2 \leq p_t \lambda_t$$

$$B_t [u_2 - p_t \lambda_t] = 0 \quad B_t \geq 0$$

and (2).

A. Unconstrained Household

For the unconstrained household (UH), (4c) holds with equality, and a positive financial bequest is left. Using (4c), (4b) can be rewritten as

$$(4b') \quad p_t w_{t+1} h' = 1.$$

This equation is used to solve for the efficient level of human capital expenditures in terms of wage rate and the market discount factor, independent of other household decisions, to obtain

$$(5a) \quad X_t = X(p_t, w_{t+1})$$

where capital letters will be used to distinguish unconstrained choices from constrained choices. The efficient human capi-

can remain constrained permanently. My argument requires families to remain either constrained or unconstrained for at least two successive generations.

³As in previous models of this type, I abstract away from life-cycle issues and sex differences. However, I shall refer to "parents" in order to simplify the exposition. One may take the last year of economic dependency to be 21 and assume that each parent has children at, say, age 30, nine years into the period of economic independence. Then, each period corresponds to roughly 30 years.

⁴As long as attention is restricted to the macroeconomic effects of a lump-sum redistribution from unconstrained to constrained families, allowing for two children of different abilities, both of whom either receive some financial bequests or none at all, would complicate but would not essentially generalize the identical-children model. Analyzing the distribution of wealth across individuals may be more interesting in this setting, but that is a topic for another paper.

tal expenditures are then subtracted from the current generation's wealth, and the efficient human capital bequest is substituted into the next generation's wealth. The utility-maximizing consumption and financial-bequest decisions are determined next by using (4a) and the current-period budget constraint. The functions describing these choices are

$$(5b) \quad C_t = C(p_t, \overset{(+)}{\widetilde{W}}_{t+1}, \overset{(+)}{\overline{W}}_t)$$

$$(5c) \quad B_t = B(p_t, \overset{(-)}{\widetilde{W}}_{t+1}, \overset{(+)}{\overline{W}}_t)$$

where

$$\widetilde{W}_{t+1} = (w_{t+1}H[X(p_t, w_{t+1})] - T_{t+1})n_{t+1}$$

is total wealth of the children inclusive of their human capital bequest and exclusive of their financial inheritance and where $\overline{W}_t = B_{t-1}/(1+g) + w_t H_t - n_{t+1}X_t - T_{t+1}$ is the total wealth of the parents exclusive of the efficient human capital expenditures. The impact of a change in p_t , a function of the interest rate, will be discussed in the general-equilibrium portion of Section II.

The unique aspect of an altruistic model of intergenerational transfers is that consumption and financial bequests will be affected by \widetilde{W}_{t+1} , the wealth of the children independent of financial inheritances. The greater is the child's wealth from other sources, the smaller is the financial bequest and the greater is parental consumption. This can be explained by noting that $\partial B_t / \partial \widetilde{W}_{t+1} = -1 + (\partial B_t / \partial \overline{W}_t)p_t$, which decomposes the total effect into a substitution and a wealth effect.⁵ An exogenous increase in the child's wealth causes B_t to fall one-for-one, a direct substitution effect. Smaller expenditures on net bequests raise the parent's wealth by p_t . Since $(\partial B_t / \partial \overline{W}_t)p_t < 1$, this

will mitigate, but not offset, the direct substitution effect.⁶

Another way of interpreting the wealth effect is from the perspective of a consolidated family-budget constraint. When the next generation's wealth exogenously increases by one unit, the present value of family wealth increases by p_t . The extent to which the increase in family wealth is passed on to the next generation is determined by $\partial B_t / \partial \overline{W}_t$.

B. Constrained Household

For the constrained household (CH), $u_2 < p_t \lambda_t$ in (4c), and with the rate of return on human capital $(w_{t+1}h' - 1)$ exceeding that on financial bequests (r_{t+1}) , it is optimal to set $B_t = 0$. There is no separation of human capital decisions and consumption in this case. Equations (4a), (4b), and (2) must be simultaneously solved for c_t , x_t , and λ_t . Solutions of this problem produce demand functions of the form

$$(6a) \quad x_t = x(w_{t+1}, \overset{(-)}{R}_{t+1}, \overset{(+)}{\overline{W}}_t)$$

$$(6b) \quad c_t = c(w_{t+1}, \overset{(+)}{R}_{t+1}, \overset{(+)}{\overline{W}}_t).$$

Due to the simultaneous nature of the solution, wealth variables now influence human capital expenditures, and the wage rate now influences consumption. A change in \overline{W}_t induces a simple wealth effect on x_t and c_t , both of which are assumed to be normal goods. Any predictable exogenous increase in the wealth of the children (R_{t+1}) will lower x_t and raise c_t . Intuitively, constrained parents share some of the children's good fortune by diverting resources toward themselves. The effect of a higher anticipated wage rate for the next generation is ambiguous and will be discussed in detail when general-equilibrium considerations are introduced in the next section.

⁵The compensated demands are found by solving the dual to the maximization problem given by (1) and (2). Equating ordinary demands to compensated demands, substituting the expenditure function to form an identity, and then differentiating gives the Hicks decomposition.

⁶Tomes (1981) and Kotlikoff (1989 Ch. 16) provide empirical support for the sign of this effect.

II. Redistribution

In this section, I discuss the effects of redistribution on capital formation. Previous studies have concentrated on the partial-equilibrium effects of a temporary redistribution of wealth. At best, this represents a starting point for a complete general-equilibrium analysis of a permanent redistribution policy. I first show that the partial-equilibrium effects of permanent and temporary policies are quite different. The general-equilibrium effects of both policies are then analyzed to assess when the partial-equilibrium impact effects may be misleading.

A. Impact Effects

The bequest savings of the unconstrained household is $S_t = p_t B_t$. Substituting (5c) into the savings expression yields a savings function in terms of the other variables in the model. Consider the direct effect of a permanent increase in lump-sum taxation on S_t . Setting $T_t = T_{t+1} = T$ and differentiating the savings function gives

$$(9) \quad \partial S_t / \partial T = -p_t \left[(\partial B_t / \partial \widetilde{W}_{t+1})(1+g) + \partial B_t / \partial \overline{W}_t \right].$$

The expression is opposite in sign and equal in absolute value to a permanent increase in the wealth of the UH. Greater wealth for the current generation will raise bequests ($\partial B_t / \partial \widetilde{W}_t$), but greater wealth for each member of the next generation will lower bequests ($\partial B_t / \partial \widetilde{W}_{t+1}[1+g]$). Since from the microeconomic analysis of the UH one knows that $\partial B_t / \partial \widetilde{W}_{t+1} = -1 + p_t \partial B_t / \partial \overline{W}_t$, an estimate of the marginal propensity to bequeath is sufficient to resolve the issue. Menchik and David (1983 p. 683) estimate $p_t \partial B_t / \partial \overline{W}_t$ to be 0.25.⁷ This implies

⁷One can also construct estimates from Tomes (1981). His table 4 is the most relevant, since there he isolates the pure effect of income changes across generations, holding constant compensating variations in transfers due to differences in income across children. The estimates of $p_t \partial B_t / \partial \overline{W}_t$ range from 0.12 to 0.16,

$\partial B_t / \partial \widetilde{W}_{t+1}(1+g) < -0.75$, suggesting that bequests fall with a rise in the permanent wealth of the family.⁸ Intuitively, given what is known about the marginal propensity to bequeath, the substitution effect of a permanent rise in family wealth is dominant, which causes parents to reduce bequests whenever they realize that the next generation is also better off.⁹

An example may help to uncover the structural factors leading to such a result. If the utility function is assumed to take the commonly employed CES form,

$$[1/(1-\sigma)]c_t^{1-\sigma} + [b/(1-\gamma)][W_{t+1}(1+g)]^{1-\gamma}$$

then (9) can be rewritten as

$$(9') \quad \frac{\partial S_t}{\partial T} = \frac{-p_t}{D} \left\{ 1 - \left(\frac{\gamma}{\sigma} \right) \left(\frac{c_t}{(\widetilde{W}_{t+1} + B_t)/(1+g)} \right) \right\}$$

where $D = p_t + (\gamma/\sigma)[c_t/(\widetilde{W}_{t+1} + B_t)] > 0$. If preferences are homothetic, $\gamma = \sigma$, then savings will increase with a permanent tax increase only if parents' consumption exceeds wealth per child, $c_t > (\widetilde{W}_{t+1} + B_t)/(1+g)$. This is certainly a possibility if there is some regression toward the mean in wealth. Consider, however, the stationary case in which wealth is equal across gen-

assuming a 40-year work life and an after-tax annual interest rate of 3 percent, as in Menchik and David (1983).

⁸Table 4 in Tomes (1981) also gives direct evidence in support of this result.

⁹This sign can be resolved more generally by writing out $\partial S_t / \partial T$, with all subscripts dropped for convenience, as

$$- \{ -[1 - (\partial B / \partial \overline{W})(1+g)/(1+r)] + \partial B / \partial \overline{W} \} \\ = (1+g)/(1+r)[1 - (\partial B / \partial \overline{W})\delta]$$

where $\delta = [(1+g) + (1+r)]/[(1+g)(1+r)] \leq 1$ if $1 \leq rg$. Taking each period to be roughly 30 years, as explained in footnote 3, with an annual population growth of 1.5 percent (each household then having approximately 1.5 children) and an annual real interest rate of 4 percent (see Edward C. Prescott, 1986), $rg = 1.26$. Thus, as long as $\partial B / \partial \overline{W} < 1$, the direct effect of a permanent lump-sum tax on the UH is to raise steady-state bequests and savings.

erations, $W_t = (\widetilde{W}_{t+1} + B_t)/(1+g) = W_{t+1}$. In this case, $c_t < (\widetilde{W}_{t+1} + B_t)/(1+g)$, because expenditures on human and financial bequests are positive. Thus, when wealth is equal across generations, preferences must be nonhomothetic ($\gamma > \sigma$) for bequests to increase with a permanent increase in taxes.

It is important to note that, unlike in Blinder's (1975) taste-bequest model, $\gamma > \sigma$ does not imply that the parental-wealth elasticity of bequests is less than 1. In an altruistic model, homothetic preferences imply a parental-wealth elasticity that is strictly greater than 1.¹⁰ Intuitively, bequests must move much more than proportionately with parents' wealth to allow the future generation's wealth to move proportionately. Therefore, the fact that bequests may fall with a permanent increase in family wealth is in no way inconsistent with the apparent empirical fact that bequests are luxury goods with respect to parents' wealth.

Now consider a permanent transfer to the constrained family. With $R_t = R_{t+1} = T$, the policy has the following impact effect on current-period human capital expenditures:

$$(10) \quad \partial x_t / \partial T = \partial x_t / \partial R_t + \partial x_t / \partial R_{t+1}.$$

As with the analysis of financial bequests, a permanent policy change produces opposing effects. However, unlike with financial bequests, there is no simple relationship between $\partial x_t / \partial R_{t+1}$ and $\partial x_t / \partial R_t$, forcing one to find estimates for *both* expressions.

Such estimates are understandably scarce, but some recent studies by Barry Chiswick and Donald Cox (1987) and Cox (1987a, b, c) are suggestive. These studies deal with the determinants of inter vivos transfers in general, rather than transfers for human capital per se. They also do not include transfers made before the recipient is 18 years of age. Nevertheless, there are some

relevant features of this work. Based on a representative cross-sectional survey of households, Cox (1987c) finds that those receiving transfers have lower current and permanent income than those not receiving transfers. Thus, the results evidently pertain to wealth-constrained families. Parents from constrained families appear to make modest transfers throughout their children's young life, rather than a larger efficient transfer early on and a financial bequest at death.

The most robust result of interest is that a permanent increase in the income of both the potential transfer recipient and his parents will increase the *likelihood* of a positive transfer.¹¹ Estimating the marginal impacts of income changes on the *level* of transfers, for those receiving transfers, is more difficult and here the empirical results are sensitive to the sample and empirical method chosen.¹² However, there is certainly no evidence to suggest that a permanent rise in family income will reduce parental transfers. If any conclusion can be drawn, it is that a permanent rise in income for the *aggregate* of wealth-constrained households will increase *aggregate* transfers by increasing the number of parents making transfers to children over 18 years old.

B. General-Equilibrium Considerations

Assuming that output is determined by a standard neoclassical production function,

¹¹ Compare the coefficients on recipient's income and the proxy for donor's income in tables 2 and 4 of Cox (1987a), column 1 in table 2 of Chiswick and Cox (1987), column a in table 5 of Cox (1987b), and tables 3, 4, and 7 of Cox (1987c). In the subsample including recipients as the secondary family unit, the studies above produce a positive, rather than negative, sign for recipient's income. The net effect is still, of course, positive.

¹² Both the sign and the statistical significance of recipient and donor income vary substantially. This may reflect the complexity of determining the factors influencing the marginal transfers from parents to children in wealth-constrained families. After the parents' altruism has been exhausted, the inefficient levels of human capital investment are likely to motivate implicit loan agreements between parents and children. The pure altruistic model does not address these types of exchanges.

¹⁰ See the discussion by Becker (1974). More precisely, one can readily couch the present model in the form of the theory from section 2 of his paper, to obtain the appropriate bequest formula from his equation 2.8.

markets are competitive, and, for convenience, the number of unconstrained and constrained households is the same implies

$$(11) \quad r_{t+1} = f'(K_{t+1}/[1+g][h_{t+1}+H_{t+1}])$$

where K_{t+1} is the capital stock per unconstrained household and f is the production function per unit of effective labor supply. This period's savings per unconstrained household provide the funds for next period's capital stock. With the past period's savings and human capital decisions taken as given at time t , equations (5) and (6) and the factor price frontier [$w_{t+1} = \varphi(r_{t+1})$, $\varphi' = -k_{t+1} = -K_{t+1}/(1+g)(h_{t+1}+H_{t+1})$] can be substituted into both (11) and the savings expression to produce

$$(12) \quad r_{t+1} = \rho(K_{t+1}, T_t, T_{t+1})$$

$$(13) \quad S_{t+1} = S(r_{t+1}, T_t, T_{t+1}).$$

In a representative-household model of bequest-savings, (12) can be sketched as a downward-sloping demand for capital, and (13) can be sketched as an upward-sloping supply of capital, holding transfers constant. However, with heterogeneous households, neither slope is restricted by theory alone. Before discussing why the slopes may become irregular and what consequences this may have for redistributive policy, it is interesting to point out that the conventional view does not necessarily follow in general equilibrium, even when the demand and supply curves have their regular slopes and the policy is temporary. An increase in T_t , holding T_{t+1} constant, shifts the supply curve leftward, as parental wealth of the UH falls, and shifts the demand curve rightward, as the increase in wealth and human capital expenditures of the CH at time t lowers the capital:labor ratio at time $t+1$, everything else held constant. This is depicted in Figure 1. The *equilibrium* level of savings may rise, fall, or remain the same depending on the relative size of the shifts. The only unambiguous effect is that real interest rates are higher and real wage rates are lower, which means that the human capital investment of the UH will be

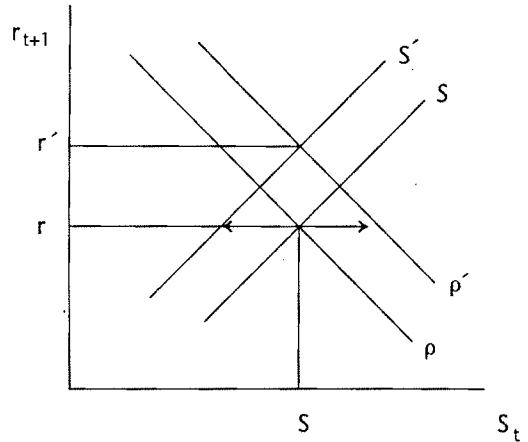


FIGURE 1. AN INCREASE IN T_t , HOLDING T_{t+1} CONSTANT, SHIFTS THE SUPPLY CURVE LEFTWARD AND SHIFTS THE DEMAND CURVE RIGHTWARD

“crowded out” by the excess demand for funds generated by the policy. Nothing can be said about the physical capital stock or the aggregate human capital stock.¹³

For a permanent policy change, even the impact effects on the demand and supply curves are ambiguous. As discussed in the previous section, the available empirical evidence suggests that a permanent increase in T will shift the supply curve and, with less certainty, the demand curve rightward. If the demand curve is downward-sloping and the supply curve upward-sloping, the level of savings will unambiguously rise. The impact on factor prices is ambiguous, and as a result it is not clear whether aggregate human capital rises or falls. However, as previously mentioned, theory does not restrict the slopes of either curve in a model with heterogeneous households.

Starting on the supply side, the substitution effect of an increase in r_{t+1} makes

¹³The general equilibrium analysis of this section is short-run in nature, since we do not examine the transition to a new steady state. Such an analysis requires the strong assumption that families stay unconstrained or constrained permanently. A steady-state analysis is carried out in a longer working paper and is available upon request.

financial bequests cheaper relative to consumption and relative to human capital investment, leading to more savings. However, an increase in r_{t+1} also induces wealth effects which are less transparent. A higher r_{t+1} raises the next generation's wealth, for a given level of financial bequests. In general equilibrium, a higher r_{t+1} means a lower w_{t+1} , causing the next generation's labor income to fall. In a representative-household model, these two wealth effects just cancel, leaving bequests unaffected. With two households, this is not the case. A higher r_{t+1} transfers wealth from the CH of the next generation, via a lower w_{t+1} , to the UH of the next generation. Since the children of the UH are made better off, parent's bequests will fall. This conflicts with the substitution effect, making the overall effect ambiguous.

On the demand side, the ambiguity arises because a change in the wage rate has an ambiguous effect on effective labor supply. A rise in w_{t+1} clearly increases the human capital investment of the UH, but not of the CH. The effect of a higher anticipated wage rate on the investment of the CH can be decomposed as

$$(10) \quad \partial x_t / \partial w_{t+1} \\ = \partial \bar{x}_t / \partial w_{t+1} + (\partial x_t / \partial R_{t+1}) h.$$

The first term is the compensated-relative-price effect of a higher return on human capital investment, and its sign is positive. The second term reflects the fact that an increase in future wages works like an exogenous increase in the child's wealth, which lowers expenditures, making the overall effect ambiguous.

Due to the possibility of irregularly sloped demand and supply curves, the general-equilibrium results may differ from the direct partial-equilibrium impacts, even after assuming that the capital market is Hicksian stable (see Peter A. Diamond, 1965; Christophe Chamley and Brian D. Wright, 1987). There are two cases in which the impact effects may be misleading. First, consider the case with a downward-sloping supply curve as depicted in Figure 2. Rightward shifts of both curves may cause a rise

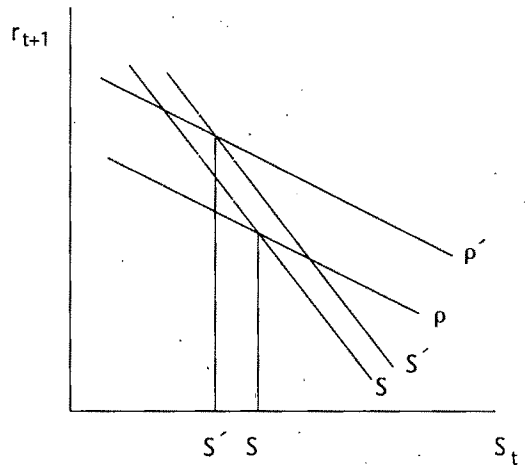


FIGURE 2. THE CASE WITH A DOWNWARD-SLOPING SUPPLY CURVE: RIGHTWARD SHIFTS OF BOTH CURVES MAY CAUSE BEQUEST-SAVINGS TO FALL IN EQUILIBRIUM

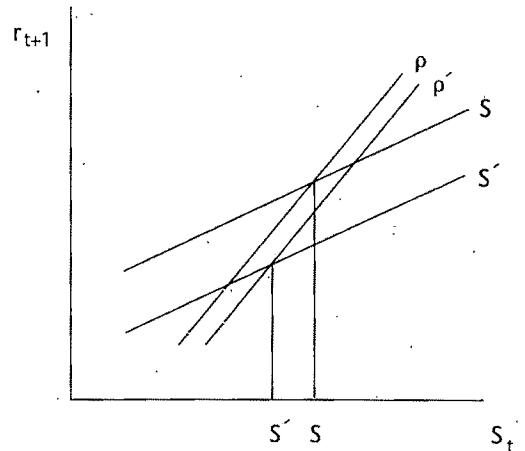


FIGURE 3. THE CASE WITH AN UPWARD-SLOPING DEMAND CURVE: RIGHTWARD SHIFTS OF BOTH CURVES MAY CAUSE BEQUEST-SAVINGS TO FALL IN EQUILIBRIUM

in interest rates which increases the wealth of the children from the UH to such an extent that bequest-savings fall in equilibrium. Second, if the demand curve is upward-sloping, rightward shifts of both curves may lower interest rates and, through a dominant substitution effect, lower savings in equilibrium as indicated in Figure 3.

III. Conclusion

Contrary to conventional belief, there is no compelling reason to believe that lump-sum redistribution will lower physical capital formation. Other authors have pointed this out for the case of life-cycle savings. The contribution of this paper is to make a similar point for the case of altruistically motivated bequests-savings. The available empirical evidence indicates that the direct effect of redistribution is to *raise* bequests for many households.

The paper also suggests that the equilibrium effects of redistribution on physical and human capital are difficult to ascertain in models with heterogeneous households, even when there is perfect knowledge concerning the impact effects. Work on a general-equilibrium simulation model with constrained and unconstrained households is currently underway. The best guess about the general-equilibrium effects of redistribution will most likely come from such a source.

REFERENCES

- Becker, Gary S., "A Theory of Social Interactions," *Journal of Political Economy*, November/December 1974, 82, 1063-93.
- , *A Treatise on the Family*. Cambridge, MA: Harvard University Press, 1981.
- and Tomes, Nigel, "Child Endowments and the Quantity and Quality of Children," *Journal of Political Economy*, August 1976, 84 (Part 2), S143-62.
- and ———, "Human Capital and the Rise and Fall of Families," *Journal of Labor Economics*, July 1986, 4, S1-39.
- Behrman, Jere R. and Taubman, Paul, "Intergenerational Earnings Mobility in the United States: Some Estimates and a Test of Becker's Intergenerational Endowment Model," *Review of Economics and Statistics*, February 1985, 67, 144-51.
- Bernheim, B. Douglas, "A Neoclassical Perspective on Budget Deficits," *Journal of Economic Perspectives*, Spring 1989, 3, 55-72.
- Blinder, Alan S., "Distribution Effects and the Aggregate Consumption Function," *Journal of Political Economy*, 1975, 83, 447-75.
- , "Intergenerational Transfers and Life Cycle Consumption," *American Economic Review*, May 1976 (*Papers and Proceedings*), 66, 87-93.
- Chamley, Christophe and Wright, Brian D., "Fiscal Incidence in an Overlapping Generations Model with a Fixed Asset," *Journal of Public Economics*, February 1987, 32, 3-24.
- Chiswick, Barry and Cox, Donald, "Inter Vivos Transfers and Human Capital Investments," Working Paper No. 103, Washington University, St. Louis, January 1987.
- Cox, Donald, (1987a) "Motives for Private Income Transfers," *Journal of Political Economy*, June 1987, 95, 508-46.
- , (1987b) "The Connection Between Public Transfers and Private Interfamily Transfers," Working Paper No. 100, Washington University, St. Louis, March 1987.
- , (1987c) "Intergenerational Transfers and Liquidity Constraints," Working Paper No. 114, Washington University, St. Louis, July 1987.
- Diamond, Peter A., "National Debt in a Neoclassical Growth Model," *American Economic Review*, December 1965, 55, 1125-50.
- Kotlikoff, Laurence, J., "Taxation and Savings: A Neoclassical Perspective," *Journal of Economic Literature*, December 1984, 22, 1576-1629.
- , "Intergenerational Transfers and Savings," *Journal of Economic Perspectives*, Spring 1988, 2, 41-58.
- , *What Determines Savings?* Cambridge, MA: MIT Press, 1989.
- Menchik, Paul L., "Intergenerational Transmission of Inequality: An Empirical Study of Wealth Mobility," *Economica*, November 1979, 46, 349-62.
- and David, Martin, "Income Distribution, Lifetime Savings and Bequests," *American Economic Review*, September 1983, 73, 672-90.
- Modigliani, Franco, "The Role of Intergenerational Transfers and Life Cycle Saving in the Accumulation of Wealth," *Journal of Economic Perspectives*, Spring 1988, 2, 15-40.

- Prescott, Edward C., "Theory Ahead of Business Cycle Measurement," *Quarterly Review* (Federal Reserve Bank of Minneapolis), Fall 1986, 10, 9-22.
- Ruhm, Christopher J., "When 'Equal Opportunity' Is Not Enough: Training Costs and Intergenerational Inequality," *Journal of Human Resources*, Spring 1988, 23, 155-72.
- Solon, Gary, Corcoran, Mary, Gordon, Roger and Laren, Deborah, "Sibling and Intergenerational Correlations in Welfare Program Participation," Working Paper No. 2334, NBER (Cambridge, MA), 1987.
- Tomes, Nigel, "The Family, Inheritance, and the Intergenerational Transmission of Inequality," *Journal of Political Economy*, October 1981, 89, 928-58.

The Welfare Economics of Price Supports in U.S. Agriculture: Comment

By GEOFF EDWARDS AND DAVID VANZETTI*

Erik Lichtenberg and David Zilberman (1986) (henceforth LZ) compare the effects of restrictions on pesticide use in United States agriculture with and without agricultural price supports. LZ's main conclusion is that price supports change the distribution of the costs of pesticide regulation. With their model specification, gains to producers from regulation in the absence of price supports became losses with a target price, while losses to consumers cum taxpayers became gains.

LZ say, and we agree, that estimating the costs of regulation "is one of economists' main contributions to the policy evaluation process" (p. 1135). A fundamental question in welfare economics is whose welfare is to be counted? LZ use a partial-equilibrium model that distinguishes welfare effects on U.S. farmers, U.S. taxpayers, and consumers. Implicitly welfare effects on consumers include those experienced by U.S. citizens and effects on consumers in the rest of the world (ROW), with the ROW effects being net of welfare changes for ROW producers. *Benefits* from pesticide regulation that arise outside the commodity market concerned (e.g., environmental benefits) are explicitly excluded from the LZ analysis, although pesticide regulation is aimed at providing such benefits.

In treating welfare impacts on foreign consumers identically to impacts on domestic consumers, LZ depart from the common practice of conducting cost-benefit analysis from a national perspective. It is of interest to ask whether the LZ results are changed if only welfare effects experienced in the

United States are counted. To help in answering this question, welfare effects of a pesticide ban that were summarized in table 2 in LZ are represented in a more disaggregated form in Table 1. The key step is the separation of changes in consumer welfare into U.S. and ROW components. This was done by taking foreigners' share in consumption of U.S. production of each commodity in the period 1980-1985 as the ROW's share in LZ's measures of consumer surplus. The ROW's shares were: corn, 26 percent; cotton, 45 percent; and rice, 51 percent (Economic Research Service, 1987). Applying these percentages to LZ's estimates of welfare losses to *world* consumers (column 2) from pesticide regulation gives welfare changes to ROW consumers as shown in column 6.¹ This simple exercise in disaggregation results in several important changes in the LZ results.

¹The elasticities of demand used by LZ in calculating welfare effects on world consumers were weighted averages of elasticities of U.S. and ROW (excess) demand. Our approach to calculating the change in ROW consumer surplus implicitly assumes that the elasticities for the two components of demand are identical. This may seem to be inconsistent with the reality that export elasticities of demand are typically much larger than domestic ones. However, simulations reveal that the size of the change in world consumer surplus and its distribution between U.S. and foreign consumers is insensitive to changes in the relationship between U.S. and ROW elasticities of demand. For corn, for example, using the weighted average elasticity of demand of -0.5 employed by LZ, the ROW share of the fall in world consumer surplus from U.S. pesticide regulations was approximately 26.0 percent when the U.S. and ROW elasticities of demand were assumed to be equal, and approximately 25.8 percent when the ROW elasticity was assumed to be 20 times as large as the U.S. elasticity. Of course, if stronger supply responses in ROW meant that the weighted average elasticities of demand for U.S. commodities were higher in the long run than the elasticities used by LZ, the ROW's share of consumer losses from U.S. pesticide regulation would fall as the United States lost export markets.

*School of Agriculture, La Trobe University, Melbourne, Australia 3083, and Department of Economics, La Trobe University, respectively. We acknowledge helpful comments by A. Myrick Freeman and an anonymous reviewer.

TABLE 1—WELFARE EFFECTS OF PESTICIDE REGULATION WITH A TARGET PRICE

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				ΔCS (WORLD)					
Commodity	ΔW (WORLD)	ΔCS (WORLD)	ΔGP (U.S.)	$-\Delta GP$ (U.S.) (2-3)	ΔPS (U.S.)	ΔCS (ROW)	ΔCS (U.S.)	$\Delta(CS - GP)$ (U.S.) (7-3)	ΔW (U.S.) (1-6)
Corn	-43	-120	-131	12	-55	-31	-89	42	-12
Cotton	-20	-88	-99	12	-32	-40	-48	51	20
Rice	-5	-25	-28	3	-8	-13	-12	16	8

Notes: ΔCS , ΔPS , ΔGP , and ΔW denote changes in consumers' surplus, producers' surplus, government payments and net welfare, respectively. All table entries are in units of \$ million.

Source: Entries in columns 1-5 are from LZ. Entries in columns 6-9 were calculated by the authors.

First, accepting LZ's estimates of welfare losses to world consumers (column 2) from pesticide regulation with a target price, losses to U.S. consumers (column 7) ranged from just under half of world consumer losses for rice to about three-quarters of world consumer losses for corn. Second, for U.S. taxpayers cum consumers, the gains (see column 8) are from $3\frac{1}{2}$ to 5 times those obtained by LZ (see column 4) for consumers (world) cum taxpayers (U.S.). Third, aggregate welfare changes for the United States with a target price (column 9) are much more favorable than the welfare changes presented by LZ for the world market (column 1). For corn, the U.S. welfare loss is less than one-third that found by LZ for the world corn market. For cotton and rice, the changes are even more dramatic: world welfare losses of \$20 million and \$5 million become welfare gains of \$20 million and \$8 million, respectively, to the United States.

It is the *domestic* welfare effects of pesticide regulation, not the global effects, that are likely to be most relevant to the U.S. debate on pesticide regulation. Disaggregation of the LZ analysis shows that, in the presence of price supports, pesticide regulation in the United States is much more attractive from a domestic perspective than from a global perspective. This point may be reinforced if (domestic) nonmarket benefits, such as environmental improvement, were incorporated.

REFERENCES

- Lichtenberg, Erik and Zilberman, David, "The Welfare Economics of Price Supports in U.S. Agriculture," *American Economic Review*, December 1986, 76, 1135-41.
- Economic Research Service (USDA), *Agricultural Outlook*, Washington, DC: U.S. Government Printing Office, March 1987.

Deloitte & Touche

Suite 2400
424 Church Street
Nashville, TN 37219-2396

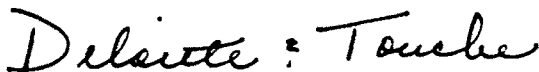
Independent Auditors' Report

Executive Committee
The American Economic Association
Nashville, Tennessee

We have audited the accompanying balance sheets of The American Economic Association as of December 31, 1990 and 1989, and the related statements of revenues and expenses, changes in general fund and restricted fund balances and cash flows for the years then ended. These financial statements are the responsibility of the Association's management. Our responsibility is to express an opinion on these financial statements based on our audits.

We conducted our audits in accordance with generally accepted auditing standards. Those standards require that we plan and perform the audit to obtain reasonable assurance about whether the financial statements are free of material misstatement. An audit includes examining, on a test basis, evidence supporting the amounts and disclosures in the financial statements. An audit also includes assessing the accounting principles used and significant estimates made by management, as well as evaluating the overall financial statement presentation. We believe that our audits provide a reasonable basis for our opinion.

In our opinion, the financial statements referred to above present fairly, in all material respects, the financial position of The American Economic Association as of December 31, 1990 and 1989, and the results of its operations and its cash flows for the years then ended, in conformity with generally accepted accounting principles.

A handwritten signature in cursive script that reads "Deloitte & Touche".

February 11, 1991

THE AMERICAN ECONOMIC ASSOCIATION BALANCE SHEETS FOR THE YEARS
ENDED DECEMBER 31, 1990 AND 1989

	Notes	1990	1989
Assets			
CASH		\$ 590,847	\$ 649,991
INVESTMENTS, at market	1, 2	4,982,272	5,756,069
ACCOUNTS RECEIVABLE, no allowance for doubtful accounts considered necessary		36,258	52,189
INVENTORY OF <i>Index of Economic Articles</i> , at cost		136,803	205,395
PREPAID EXPENSES		29,288	31,448
OFFICE FURNITURE AND EQUIPMENT—at cost, less accumulated depreciation of \$94,592 (1990) and \$85,151 (1989)		68,509	59,774
TOTAL		<u>\$5,843,977</u>	<u>\$6,754,866</u>
Liabilities and Fund Balances			
ACCOUNTS PAYABLE AND ACCRUED LIABILITIES		\$ 452,139	\$ 600,843
DEFERRED REVENUE:	1		
Membership dues		752,148	749,293
Subscriptions		486,647	517,102
<i>Job Openings for Economists</i>		26,955	24,241
Total deferred revenue		1,265,750	1,290,636
ACCRUAL FOR SURVEY OF MEMBERS	1	224,853	159,051
FUND BALANCES:			
General		3,929,790	4,641,042
Net worth		3,929,790	4,641,042
Restricted	1	(28,555)	63,294
Total fund balances		3,901,235	4,704,336
TOTAL		<u>\$5,843,977</u>	<u>\$6,754,866</u>

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF REVENUES AND EXPENSES
FOR THE YEARS ENDED DECEMBER 31, 1990 AND 1989

	Notes	1990	1989
REVENUES FROM DUES AND ACTIVITIES:			
Membership dues and subscriptions		\$1,113,113	\$1,048,752
Nonmember subscriptions		775,803	759,282
<i>Job Openings for Economists</i> subscriptions		38,551	37,498
Advertising		142,274	134,428
Sale of <i>Index of Economic Articles</i>		206,752	90,695
Sale of copies, republications, and handbooks		44,579	39,780
Sale of mailing list		64,322	56,591
Annual meeting		82,286	83,426
Sundry (Exhibit I)		123,463	83,296
Operating revenues		2,591,143	2,333,748
PUBLICATION EXPENSES:			
<i>American Economic Review</i>		802,272	743,755
<i>Journal of Economic Literature</i>		1,023,621	950,835
<i>Journal of Economic Perspectives</i>		431,563	393,262
<i>Job Openings for Economists</i>		60,269	62,440
<i>Survey of Members</i>	1	70,000	70,000
<i>Index of Economic Articles</i>		227,148	50,272
		2,614,873	2,270,564
OPERATING AND ADMINISTRATIVE EXPENSES:			
General and administrative:			
Salaries		221,748	205,473
Rent		23,309	22,001
Other (Exhibit II)		235,992	214,814
Committee		64,767	63,894
Annual meeting		11,797	6,608
		557,613	512,790
Operating expenses		3,172,486	2,783,354
Operating deficit		(581,343)	(449,606)
INVESTMENT INCOME RECOGNIZED	2	320,303	297,054
EXPENSES IN EXCESS OF REVENUES		(\$ 261,040)	(\$ 152,552)

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN GENERAL FUND BALANCE
FOR THE YEARS ENDED DECEMBER 31, 1990 AND 1989

	Total	Operations	Market Value Adjustments
Balance at December 31, 1988	\$3,966,108	\$2,187,599	\$1,778,509
Change in market value of investments	827,486	—	827,486
Expenses in excess of revenues	(152,552)	(152,552)	—
Balance at December 31, 1989	4,641,042	2,035,047	2,605,995
Change in market value of investments	(450,212)	—	(450,212)
Expenses in excess of revenues	(261,040)	(261,040)	—
Balance at December 31, 1990	\$3,929,790	\$1,774,007	\$2,155,783

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN RESTRICTED FUND BALANCE FOR THE
YEARS ENDED DECEMBER 31, 1990 AND 1989

	Balance at January 1	Receipts	Disburse- ments	Balance at December 31
YEAR ENDED DECEMBER 31, 1989:				
The Alfred P. Sloan Foundation, Ford Foundation, Federal Reserve System and Rockefeller Foundation grants for the increase of educational opportunities for minority students in economics	\$188,838	\$115,136	\$235,786	\$68,188
The Andrew W. Mellon Foundation, Alfred P. Sloan Foundation and National Science Foundation grants to study economic graduate education in the United States	115,960	178,886	317,322	(22,476)
The Olin Foundation grant to confer on intellectual property rights	—	12,500	3,011	9,489
The Minority Scholarship Fund for minority students applying for graduate work in economics	5,000	—	—	5,000
Sundry	8,493	1,100	6,500	3,093
	<u>\$318,291</u>	<u>\$307,622</u>	<u>\$562,619</u>	<u>\$ 63,294</u>
YEAR ENDED DECEMBER 31, 1990:				
The Alfred P. Sloan Foundation, Ford Foundation and Federal Reserve System grants for the increase of educational opportunities for minority students in economics	\$ 68,188	\$246,070	\$361,559	(\$ 47,301)
The Andrew W. Mellon Foundation, Alfred P. Sloan Foundation and National Science Foundation grants to study economic graduate education in the United States	(22,476)	161,114	132,481	6,157
The Olin Foundation grant to confer on intellectual property rights	9,489	—	5,093	4,396
The Minority Scholarship Fund for minority students applying for graduate work in economics	5,000	—	—	5,000
Sundry	3,093	100	—	3,193
	<u>\$ 63,294</u>	<u>\$407,284</u>	<u>\$499,133</u>	<u>(\$ 28,555)</u>

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CASH FLOWS
FOR THE YEARS ENDED DECEMBER 31, 1990 AND 1989

	1990	1989
CASH FLOWS FROM OPERATING ACTIVITIES:		
Receipt of membership dues and subscriptions	\$2,582,188	\$2,468,136
Disbursements to suppliers and employees	(3,163,068)	(2,853,839)
Receipts of restricted funds	407,284	307,622
Disbursements made from restricted funds	(499,133)	(562,619)
NET CASH USED IN OPERATING ACTIVITIES	(672,729)	(640,700)
CASH FLOWS FROM INVESTING ACTIVITIES:		
Purchase of office furniture and equipment	(30,303)	(18,856)
Purchases of investments	(1,685,393)	(2,951,478)
Proceeds from sale of investments	1,958,288	2,724,014
Proceeds from interest and dividends on investments	370,993	564,739
NET CASH PROVIDED BY INVESTING ACTIVITIES	613,585	318,419
NET DECREASE IN CASH	(59,144)	(322,281)
Cash at Beginning of Year	649,991	972,272
Cash at End of Year	<u>\$ 590,847</u>	<u>\$ 649,991</u>
RECONCILIATION OF EXPENSES IN EXCESS OF REVENUES TO NET CASH USED IN OPERATING ACTIVITIES:		
Expenses in excess of revenues	(\$ 261,040)	(\$ 152,552)
Adjustments to reconcile expenses in excess of revenues to net cash used in operating activities:		
Depreciation	21,568	21,318
Changes in assets, liabilities, and fund balances:		
Decrease in accounts receivable	15,931	101,957
(Increase) decrease in inventory	58,592	(26,736)
Decrease (increase) in prepaid expenses	2,160	(8,775)
(Decrease) increase in accounts payable and accrued liabilities	(148,704)	62,575
(Decrease) increase in deferred revenue	(24,886)	32,431
Increase (decrease) in accrual for <i>Survey of Members</i>	65,802	(118,867)
Investment income recognized	(320,303)	(297,054)
Decrease in restricted fund balance	(91,849)	(254,997)
Total adjustments	<u>(411,689)</u>	<u>(488,148)</u>
NET CASH USED IN OPERATING ACTIVITIES	(\$ 672,729)	(\$ 640,700)
SUPPLEMENTAL SCHEDULE OF NONCASH INVESTING TRANSACTIONS:		
The Association transferred investment earnings net of investment income to the general fund:		
Investments	\$ 450,212)	(\$ 827,486)
General fund	(450,212)	827,486
	<u>\$ —</u>	<u>—</u>

See notes to financial statements.

The American Economic Association Notes to Financial Statements for the Years Ended December 31, 1990 and 1989

1. Summary of Significant Accounting Policies

Investments are accounted for on a market-value basis. The investment income recognized is modified to reflect only the Association's approximate historical average rate of return, which is currently 5 percent. Investment income represents 5 percent of the total cash and market value of investments at the beginning of the year. The change in market value of investments and dividends and interest earned net of investment income recognized is recorded directly to the general fund.

Accrual for Survey of Members. Every three to five years, the Association publishes a survey which lists, among other things, the names and addresses of its membership. This survey was most recently published in 1989 and distributed at no cost to the membership. In order to properly match the publishing cost of this survey with revenue from membership dues, the Association provided \$70,000 in 1990 and 1989 for estimated publishing costs which will reduce actual survey expenses in the year of publication.

Deferred revenue represents income from membership dues and subscriptions to the various periodicals of the Association which are deferred when received. These amounts are then recognized as income following the distribution of the specified publications to the members and subscribers of the Association. Income from life membership dues is recognized over the estimated average life of these members.

The American Economic Association files its federal income tax return as an educational organization, substantially exempt from income tax under Section 501(c)(3) of the Internal Revenue Code. As required by Section 511(a) of this Code, the Association provides for federal income taxes on certain revenues which are not substantially related to its tax-exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing lists. The Association has been determined to be an organization which is not a private foundation.

Certain restricted funds are administered on a reimbursement basis; therefore, disbursements are allowed prior to receipt of grant proceeds.

Certain reclassifications have been made to the 1989 amounts in order to conform to the 1990 presentation.

2. Investments and Investment Income

Investments consist of:

	December 31, 1990		December 31, 1989	
	Cost	Market	Cost	Market
Government obligations, bonds and commercial paper	\$ 627,144	\$ 645,520	\$ 586,999	\$ 617,179
Mutual funds	<u>4,170,729</u>	<u>4,336,752</u>	<u>4,578,432</u>	<u>5,138,890</u>
	<u>\$4,797,873</u>	<u>\$4,982,272</u>	<u>\$5,165,431</u>	<u>\$5,756,069</u>

Investment income recognized consists of:

	Year Ended December 31	
	1990	1989
Government obligations, bonds and commercial paper—interest	\$ 70,436	\$ 92,817
Corporate stocks and mutual funds—cash dividends	300,557	471,922
Corporate stocks and mutual funds—net gain (loss) on sale	(94,663)	11,094
Change in market value	(406,239)	548,707
Transfer to general fund, net	<u>450,212</u>	<u>(827,486)</u>
Investment income recognized, net	<u>\$320,303</u>	<u>\$297,054</u>

3. Retirement Annuity Plan

Employees of the Association are eligible for participation in a contributory retirement annuity plan. Payments by the Association and participating employees are based on the employee's compensation. Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was approximately \$50,000 and \$42,000 for 1990 and 1989, respectively.

4. Ratio of Net Worth to Expenses

The ratio of net worth at December 31, 1990 to 1991 budgeted expenses is 1.24 and the ratio of net worth at December 31, 1989 to actual 1990 expenses is 1.47.

EXHIBIT I—THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF SUNDRY REVENUES FOR THE YEARS ENDED DECEMBER 31, 1990 AND 1989

	1990	1989
AER submission fees	\$49,708	\$49,855
Royalties— <i>Dialog</i>	22,917	20,025
Royalties— <i>Silver Platter</i>	10,316	—
Royalties—other	9,743	7,808
CSWEP membership dues	29,243	4,721
Donations	1,027	368
Permission to reprint	329	327
Foreign postage	106	192
Miscellaneous income	74	—
	<u>\$123,463</u>	<u>\$83,296</u>

EXHIBIT II—THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF OTHER GENERAL AND ADMINISTRATIVE EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1990 AND 1989

	1990	1989
Dues and subscriptions	\$ 60,582	\$ 59,183
Mailing list file maintenance	30,408	33,446
Postage	26,265	26,501
Depreciation (straight-line method)	21,568	21,318
Accounting and legal	22,211	20,292
Investment counsel and custodian fees	16,343	16,801
Office supplies	20,526	11,547
President and president-elect expenses	6,549	7,227
Insurance and miscellaneous	10,551	6,722
Election expenses	12,275	5,936
Telephone	5,277	4,479
Bank charges	3,158	713
Travel	279	649
	<u>\$235,992</u>	<u>\$214,814</u>

Begin making plans to attend the

Annual Meeting of The American Economic Association

(in Conjunction with Allied Social Science Associations)
to be held in

NEW ORLEANS, LA

Jan. 3-5, 1992

The Employment Center opens Thursday, January 2; sessions begin Friday, January 3.

See the September *AER* for the American Economic Association's preliminary program.

The 1993 meeting will be held in Anaheim, CA, January 5-7, 1993.



The Governmental Habit Redux **Economic Controls from Colonial Times to the Present**

Jonathan R. T. Hughes

To the distinguished economic historian Jonathan Hughes, the ambiguous outcomes of attempted deregulation signal America's urgent need to probe the origins of our vast and chaotic maze of government economic controls. Why do government restrictions on the economy continue to proliferate, in spite of avowed efforts to allow the market a freer rein? How did this complicated network of nonmarket economic controls come about and whose purposes does it serve? While exploring these questions, Hughes updates his classic book *The Governmental Habit* to reflect the experience of what he calls the "wild ride" of the last fifteen years.

"Jonathan Hughes is probably the only living economic historian with the breadth of vision and open-mindedness to produce a book like this."—Hugh Rockoff

Cloth: \$29.95 ISBN 0-691-04272-1

The Business Cycle **Growth and Crisis under Capitalism**

Howard J. Sherman

Are the recurring recessions of the capitalist world merely short-term adjustments to changing economic circumstances in a system that tends, in general, toward equilibrium? In this accessible study of the business cycle, Howard Sherman makes a powerful case that recessions and painful involuntary unemployment are endogenous to capitalism. Drawing especially on the work of Wesley Clair Mitchell, Karl Marx, and John M. Keynes, Sherman explains why the nature of the business cycle produces serious economic loss and misery during its contraction phase, just as it produces growth in its expansion phase. For anyone interested in how the U.S. economy operates, this book will be an invaluable resource.

"This book will not only contribute to the existing literature on business cycles but will also become a standard reference for all students and researchers in the field."—Martin H. Wolfson

Cloth: \$45.00 ISBN 0-691-04262-4

The World Trading System at Risk

Jagdish Bhagwati

Jagdish Bhagwati, one of the world's leading economists, offers a fascinating overview of the perils and promise facing the world trading system. He refutes facile but fashionable criticisms of the General Agreement on Tariffs and Trade (GATT). Warning of the dangers of flouting the GATT's provisions, he shows that its underlying conception of trading by rules will be undermined if the U.S. extends accusations of "unfair trade" practices to areas as diverse as retail distribution systems, infrastructure spending, saving rates, and workers' rights.

"... a thoughtful, scholarly, and authoritative analysis of the current status and future potential of GATT."—Milton Friedman

Cloth: \$16.95 ISBN 0-691-04284-5

Not available from Princeton in the United Kingdom and Europe

Efficiency Wages

Models of Unemployment, Layoffs, and Wage Dispersion

Andrew Weiss

Known for his seminal work in efficiency-wage theory, Andrew Weiss surveys recent research in the field and presents new results. He shows how wage schedules affect the kinds of workers a firm employs and how well those workers perform on the job. Using straightforward examples, he demonstrates how efficiency-wage theory can explain labor market outcomes and guide government policy. There is a separate section of applications to less developed countries.

"Efficiency Wages should be on the bookshelf of all labor and macroeconomists."

—Lawrence H. Summers, *Harvard University*

Paper: \$8.95 ISBN 0-691-00388-2 Cloth: \$29.50 ISBN 0-691-04279-9

Not available from Princeton in the United Kingdom and Europe

Essays on the Intellectual History of Economics

Jacob Viner

Edited by Douglas A. Irwin

Ranking among the most distinguished economists and scholars of his generation, Jacob Viner is best remembered for his work in international economics and in the history of economic thought. Never before, however, have Viner's important contributions to the intellectual history of economics been collected into one convenient volume.

"Jacob Viner was a great and original economic theorist. What is rarer, Viner was a learned scholar. What is still rarer, Viner was a wise scientist. This new anthology of his writings on intellectual history is worth having in every economist's library—to sample at intervals over the years in the reasoned hope that Viner's wisdom will rub off on the reader and for the pleasure of his writing."—Paul A. Samuelson

Cloth: \$49.50 ISBN 0-691-04266-7

New in paperback

Crisis amid Plenty

The Politics of Soviet Energy under Brezhnev and Gorbachev

Thane Gustafson

Although the Soviet Union has the most abundant energy reserves of any country, energy policy has been the single most disruptive factor in its industry since the mid-1970s. This major case study treats the paradox of the energy crisis as an essential part of larger economic problems of the Soviet Union and as a key issue in determining the fate of the Gorbachev reforms.

"... should interest Soviet specialists, in both universities and government, and prove instructive for students of the Soviet economic system, graduate or undergraduate."—Choice

A Rand Corporation Research Study

Now in paper: \$14.95 ISBN 0-691-02340-9



PRINCETON UNIVERSITY PRESS

41 WILLIAM ST. • PRINCETON, NJ 08540 • (609) 258-4900

ORDERS: 800-PRS-ISBN (777-4726) • OR FROM YOUR LOCAL BOOKSTORE

ANNOUNCING

TSP VERSION 4.2

Enhancements in
Version 4.2 include:

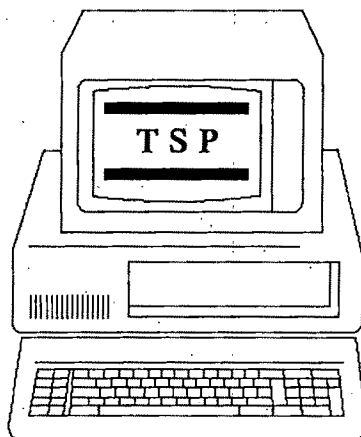
NEW TIME SERIES PROCEDURES

- Vector autoregressions
- Kalman filter estimation
- Shiller lags
- GARCH-M estimation
- Dickey-Fuller distribution for testing

OTHER NEW FEATURES

- Panel data estimation
- GMM estimation
- Matrix algebra with easy expressions like $(X'X)^{-1}X'Y$

... and much more.



**Now the oldest econometric software
package is better than ever.**

If you have a DEC Vax, IBM mainframe, PC/286/386 or compatible, or Unix workstation, then we have a copy of Version 4.2 for you.

**ALSO AVAILABLE NOW:
TSP VERSION 4.1B FOR MACINTOSH**

For more information, contact us at

TSP International

P.O. Box 61015, Station A · Palo Alto, CA 94306

Phone (415) 326-1927 · FAX (415) 328-4163

Bitnet: TSPINTL@SUWATSON

1991 ANNUAL MEMBERSHIP RATES

Membership includes a subscription to the *American Economic Review* (quarterly) plus *Papers and Proceedings*, the *Journal of Economic Literature* (quarterly) and the *Journal of Economic Perspectives* (quarterly). You may elect to receive two of the three journals; dues will be reduced by *\$6.00 per year. Note: Regardless of your selection, you will receive the *Papers and Proceedings*.

JOURNAL OPTIONS AVAILABLE AT ANNUAL RENEWAL ONLY. Single copies may be purchased for \$15.00 each after publication; prepayment required.

- Regular members with annual incomes of \$30,000 or less \$44.00
- Regular members with annual incomes above \$30,000 but no more than \$40,000 \$52.80
- Regular members with annual incomes above \$40,000 \$61.60
- Junior members [available to registered students for five years maximum

(eff. 1/1/90)]. Student status must be certified yearly by your major professor or school registrar \$22.00

- In Countries other than the U.S.A., Add \$16.00 to cover postage.
- Family members (persons living at the same address as a regular member, additional memberships without subscription to the publications of the Association) \$8.80

Please begin my membership in _____ (year) for the following period:

☐ Jan.-Dec. ☐ April-March ☐ July-June ☐ Oct.-Sept.

First Name and Initial	Last Name	Suffix
Address Line 1		
Address Line 2		
Address Line 3		
City		
State or Country	Zip/Postal Code	

MAJOR FIELDS (TWO ONLY)
 LIST FIELDS WITH WHICH YOU CURRENTLY IDENTIFY. SELECT FIELD CODE FROM JEL, "Classification System for Books".

Please type or print information above. Please pay with a check or money order payable in United States Dollars. **Canadian and foreign payments must be in the form of a check drawn on a United States bank payable in United States Dollars.** Please note: It is the policy of the Association, not to refund membership payments.

Endorsed by (AEA member) _____

Below for Junior Members Only

I certify that the person named above is enrolled as a student at _____

You will receive all three Association Publications unless indicated below. I do NOT want to receive the following publication. Select only one.

- (1) AER ☐
 (2) JEL ☐
 (3) JEP ☐

Authorized Signature _____

Please include current telephone and fax numbers in space provided.

Office _____
 Home _____
 Fax _____

Dues (includes AER, JEL & JEP) \$ _____

*Less \$6.00 credit if applicable. (_____) (I understand I will NOT receive the publication selected above.)

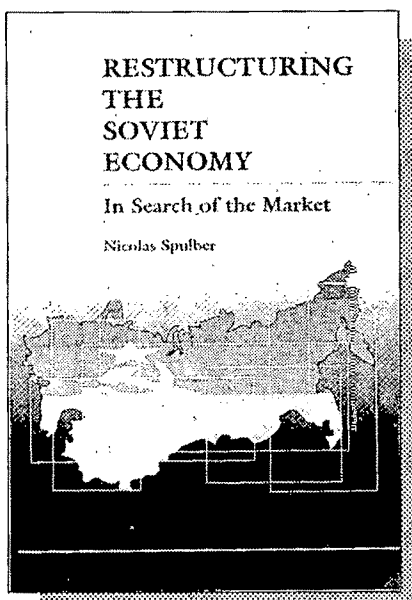
PLEASE SEND WITH PAYMENT TO:

**American Economic Association
 2014 Broadway, Suite 305
 Nashville, Tennessee 37203
 U.S.A.**

Optional donation. _____

Enter amount enclosed. \$ _____ USD

MICHIGAN MICHIGAN MICHIGAN



Nicolas Spulber

RESTRUCTURING THE SOVIET ECONOMY

In Search of the Market

cloth \$39.50

Kenneth Koford and Jeffrey Miller,
Editors

SOCIAL NORMS AND ECONOMIC INSTITUTIONS

cloth \$37.50

James M. Buchanan

THE ECONOMICS AND THE ETHICS OF CONSTITUTIONAL ORDER

cloth \$39.50

Thomas R. Palfrey, Editor

LABORATORY RESEARCH IN POLITICAL ECONOMY

cloth \$42.50

Raymond M. Duch

PRIVATIZING THE ECONOMY

Telecommunications Policy in
Comparative Perspective

cloth \$34.50

Zoltan J. Acs and David B. Audretsch,
Editors

INNOVATION AND TECHNOLOGICAL CHANGE

An International Comparison

cloth \$42.50

Akira Takayama

ANALYTICAL METHODS IN ECONOMICS

cloth \$54.50 / paper \$24.95



THE UNIVERSITY OF MICHIGAN PRESS

Dept. NP Ann Arbor, Michigan 48106-1104

ECONOMICS

NEW FROM CALIFORNIA

Panama at the Crossroads

Economic Development and Political Change in the Twentieth Century

ANDREW ZIMBALIST & JOHN WEEKS

"This book explains why Noriega was largely irrelevant to the issues that plague U.S.-Panama relations, and why deep-seated economic and social problems threaten to tear apart both Panama and U.S. policies in the region."

—Walter LaFeber, Cornell University
256 pages, \$40.00 cloth, \$13.95 paper

Between Feminism and Labor

The Significance of the Comparable Worth Movement

LINDA M. BLUM

"Blum develops an astute and groundbreaking analysis of the comparable worth strategy for gender pay equity."

—Judith Stacey, author of
Patriarchy and Socialist Revolution in China
259 pages, \$30.00 cloth, \$11.95 paper

Shoshaman

A Tale of Corporate Japan

ARAI SHINYA

Translated by Chieko Mulhern

"Arai's novel gives us an absorbing, realistic glimpse inside a Japanese trading company."

—Daniel I. Okimoto,
author of *Between MITI and the Market*
Voices from Asia
248 pages, \$35.00 cloth, \$12.95 paper

Japan's Administrative Elite

B. C. KOH

Now in paper—"A thorough and intelligent account of what is undoubtedly the main-spring of Japan's uniquely effective economic system."—*Far Eastern Economic Review*
312 pages, 2 figures, 38 tables, \$12.95 paper

Historical Economics

Art or Science?

CHARLES P. KINDLEBERGER

Drawing on history, politics, cultural anthropology, sociology, and geography, historical economics bridges the gap between abstraction and fact engendered by traditional economic science. These essays cover a range of historical periods.

381 pages, \$34.95 cloth

Regulatory Choices

A Perspective on Developments in Energy Policy

Edited by RICHARD J. GILBERT

This is the first comprehensive economic history of energy policy and its consequences in California, where some of the most innovative programs of regulatory reform have originated.

416 pages, 114 tables and graphs, \$39.95 cloth

Strategies for Learning

Small-Group Activities in American, Japanese, and Swedish Industry

ROBERT E. COLE

Now in paper—"This is must reading for scholars. Cole shows how to organize knowledge through the use of models that provide meaningful patterns and insight."

—*Academy of Management Review*
364 pages, 6 figures, 5 tables, \$13.95 paper

AT BOOKSTORES OR ORDER TOLL-FREE
1-800-822-6657. VISA/MASTERCARD.

UNIVERSITY OF CALIFORNIA PRESS
BERKELEY LOS ANGELES NEW YORK OXFORD

Announcing a New Firm Specializing in Finance

Kolb Publishing Company

Understanding Futures Markets (3rd Edition)—Robert W. Kolb

The third edition of this leading text maintains the high readability and teachability of the first two editions. The new edition extends the coverage of financial futures and includes many more examples. It has the most up-to-date coverage of hot topics such as Globex, the accelerating internationalization of the markets, and the continuing struggle between the CFTC and the SEC. Now with an instructor's manual. Available January 2, 1991.

Options: An Introduction—Robert W. Kolb

This completely new text provides a comprehensive introduction to options. After discussing the market environment, the text analyzes option values at expiration and the no-arbitrage bounds on option prices before expiration. A thorough discussion of the Binomial Model provides a stepping stone to the Black-Scholes Model. The book also covers currency options, stock index options, and options on futures. Included with the text is **OPTION!** software. With **OPTION!**, students can analyze and graph profits and losses for option combinations; analyze binomial option values with multiple periods; find Black-Scholes option values; approximate Black-Scholes values with the binomial model; and simulate stock and option price paths. The software is fully integrated into the text to provide a comprehensive teaching package. Instructor's manual. Available now.

The International Finance Reader—edited by Robert W. Kolb

This text provides the most timely, comprehensive, and readable selection of articles on international finance. The text includes a mix of enduring classics and the best articles on recent trends in global finance and banking. It includes coverage of the Japanese financial markets, the debt crisis of the developing countries, and the international financial implications of Eastern Europe's emergence from Communism. Available now.

Financial Institutions and Markets: A Reader—edited by Robert W. Kolb

By focusing on current and readable articles, this text stays apace of the continuing financial revolution in the United States and abroad. Designed as a course supplement, the text contains very recent and highly readable articles to bring the real world into the classroom. Topics include: the savings and loan crisis, the changing face of U.S. financial regulation, new financial instruments, the interaction of stocks, futures and options, the junk bond market, and innovations in the management of financial institution. Available now.



Kolb Publishing Company
11355 S.W. 67th Avenue Miami, Florida 33156
(305) 663-0550 FAX (305) 663-6579



AEA Life Insurance Is as Individual as You Are.

Our Term Life Insurance Plan is custom-designed for members of our profession. What's more, each policy can then be tailored to suit your individual needs. As these needs change, so can the policy—and it can stay with you no matter how often you change jobs.

Our group purchasing power helped us to negotiate top quality insurance, at a very low price. To take advantage of this benefit of

membership, call 1-800-424-9883 for further details (in Washington, D.C. call 457-6820).

AEA INSURANCE
Designed by Members.
For Members.

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

QUESTION: Where is the "best" place to reach over 15,000 professionals who need your expert services?

ANSWER: The Expert Services Section of **Best's Directory of Recommended Insurance Attorneys!** Since 1928, we have addressed the needs of insurance defense attorneys and other legal/claims officers by providing this "one-of-a-kind" publication.

Our **Expert Services Section** identifies professionals providing assistance to the insurance and legal defense industries. Each firm listing contains information on services rendered, membership affiliations and firm members. Here are just some of the services represented in this section:

- Independent Shorthand Reporters
- Special Investigators
- Structured Settlement Firms
- Testing Laboratories
- Handwriting Experts



- Engineering Experts
- Environmental Specialists
- Meteorologists
- Rehabilitation
- Construction Consultants
- Economic Consultants
- Consultants

Having your firm included in this publication is truly a mark of distinction. For over 90 years, **A.M. Best** publications have been regarded as *authoritative, reliable and accurate*, because we always verify our information before releasing it to the public.

A listing in this Directory begins by filing a "Confidential Firm Report." Each individual or firm that applies for a listing is required to provide a list of clients for verification and recommendation. Upon acceptance of these credentials, a nominal fee will be charged to be listed in this prestigious publication.

Don't miss this exciting opportunity — **ACT NOW** by completing the attached coupon and mail, or fax the information to (908) 439-3296 to begin the selection process. Please respond immediately, as we need this information *no later than August 1, 1991*.

____ Yes, I would like further information on the Expert Services Section of Best's Directory of Recommended Insurance Attorneys.*

Name _____

Title _____

Company _____

Type of Service _____

Street Address _____

City _____ State _____ Zip _____

Signature _____ Date _____

Telephone _____



The Insurance Information Source™

A.M. Best Company

Ambest Road, Oldwick, NJ 08858-9988

Attention: Chrystine Koehler

Legal/Claims Department

(908) 439-2200 extension 4531

Fax: (908) 439-3296

* For additional information and questions, please call Chrystine Koehler, Assistant Mgr., Legal/Claims Department.

1165

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

INTERNATIONAL ECONOMIC JOURNAL

VOLUME 5

SPRING 1991

NUMBER 1

- | | |
|--|--|
| Debt Buybacks and Forgiveness in a Model with Voluntary Repudiation | Peter B. Kenen |
| Export Diversification and Export Instability: Some Evidence
from South East Asia and Latin America | A. Savvides and H. Mohtadi |
| Testing Alternative Specifications of Reserve Flow Equations:
The Japanese Experience, 1959-1986 | Bun Song Lee and Mark E. Wohar |
| Import Contents of Final Expenditures in the United States | C. S. Yan and E. C. Koziara |
| Effects of Exchange Controls on Black Market Exchange Rate
in a Less Developed Country | K. Gyimah-Brempong |
| Specification Tests of the Aggregate Import Demand Model
in Developing Countries | Augustine Arize |
| An Analysis of Optimality of Housing Investment in Korea | K. H. Kim and S. H. Suh |
| Economic Growth and the Expanding Export Sector:
China 1952-1985 | Andy C. C. Kwan and John A. Cotsomitis |

MANUSCRIPTS SUBMITTED (in triplicate) in Asia for publication should be sent to the Managing Editor, Professor Wontack Hong, Department of International Economics, College of Social Sciences, Seoul University, Seoul 151-742, Korea. Manuscripts from countries outside Asia should be sent to the Co-Editor, Professor Young Chin Kim, Department of Economics, Northern Illinois University, DeKalb, Illinois 60115, U. S. A. The personal subscription price of a volume (which includes postage) is \$40.00 per year. The institutional subscription price is \$50.00 per year. Cheques should be made payable to the International Economic Journal and sent directly to the Managing Editor. (Volumes 1, 2, 3 & 4 are available.)

eDATA version 2.91

Companion Program/Data Disks for the Economic Report of the President

eDATA - a general purpose economic database constructed from ALL statistical tables of the Economic Report of the President and more - is designed to help you with data access and presentation easily. eDATA not only contains large amount of economic and business data, but also provides efficient data retrieval, index search, display, graph, print, transform, analyze and transfer of data series.

You can run eDATA to retrieve a dozen data series, print or graph them in slides for presentation. You can even transfer the retrieved or transformed data to Lotus 1-2-3, dBASE III/IV, or an econometric program and estimate a regression based on the transferred data, all in a few minutes. A student version of eDATA containing 1/10 of database is also available.

System Requirements: IBM PC/XT/AT/PS2 and compatibles; MS-DOS 2.0 or later; 384K memory.

Name: _____
Affiliation: _____
Address: _____

Phone: () _____ - _____

_____ copies@ US\$99.50
_____ updates@ US\$59.50
_____ student version @ US\$29.50

free first class shipping in U.S. and Canada; other international air mail add \$5; for updates please send original program and data diskettes of an older version.

Diskette: ☐ 51/4 ☐ 31/2

Payment Enclosed: US\$ _____
personal check, money order, university or government purchase order.

To order, complete the above coupon and send to ▼

APPLIED DATA ASSOCIATES P. O. Box 8976, University Station, Portland, OR 97207

PUBLIC GOODS, MIXED GOODS, AND MONOPOLISTIC COMPETITION

Stephen Shmanske

IN THIS BOOK, Stephen Shmanske builds a theoretical bridge between public goods and monopolistic competition, suggesting that they are different dimensional simplifications of the same general model. His conclusions have great relevance to policy formation on public goods provision.

Shmanske argues that public goods models have usually ignored or simplified the utilization dimension. Furthermore, private goods models in the monopolistic competition vein have two implicit dimensions of consumption, but again, one is treated in a very constrained fashion. The general, mixed goods model draws from both traditions, using the results of one model to generalize and extend the other. 200 pp. 9 b&w illus. Bib. Index. \$45.00s

Available from:



TEXAS A&M UNIVERSITY PRESS

Drawer C ♦ College Station, Texas 77843-4354

Order toll-free: 1-800-826-8911 ♦ fax: 409-847-8752

A History of Interest Rates

Sidney Homer and Richard Sylla

New Third Edition

With a foreword by Henry Kaufman

Sylla has revised and updated this third edition, which now carries the history of interest rates to 1989. Highlights of the new edition include documentation of the recent ascent of interest rates throughout the world to their highest levels since the Middle Ages, a great deal of

new information on Japan and the European Economic Community, more information on real interest rates and consumer interest rates, expanded treatment of the United States, and updated charts and tables.

1991. 675 pp. Cloth, \$75.00

The Power of the Financial Press:

Journalism and Economic Opinion in Britain and America

Wayne Parsons

"The story of the key role the press played in reversing the economic policy of 50 years is well told." —*Harvard Business Review*

1990. 266 pp. Cloth, \$24.95

Rutgers University Press, 109 Church Street, New Brunswick, NJ 08901

Cambridge University Press

The Return of Scarcity

Strategies for an Economic Future

H. C. Coombs

The distinguished Australian economist's essays on resource use, allocation, and the environment, challenge current policy maker assumptions by arguing for an economic system to better serve social responsibilities.

36373-X Hardcover \$39.95

36896-0 Paper \$17.95

Economic Forecasting:

An Introduction

Ken Holden, David Peel and John Thompson

This work is the only currently available text that provides comprehensive coverage of the methods and applications in the rapidly developing field of forecasting the future state of the economy.

35612-1 Hardcover about \$49.50

35692-X Paper about \$16.95

The Stockholm School of Economics Revisited

Lars Jonung, Editor

In this volume leading scholars look at the heritage and impact of the important work done by the Stockholm School from the 1920s to the present.

Historical Perspectives on Modern Economics

39127-X Hardcover \$64.50

Leading Economic Indicators

New Approaches and
Forecasting Records

**Kajal Lahiri and Geoffrey H. Moore,
Editors**

This volume comprises articles on new approaches to indicator research, and covers advances in three areas of research: the use of new developments in economic theory and time-series analysis to rationalize existing systems of indicators; more appropriate methods to evaluate the forecasting records of leading indicators; and the development of new indicators.

37155-4 Hardcover \$59.50

Mismatch and Labour Mobility

Fiorella Padoa Schioppa, Editor

This conference proceedings volume examines the evidence on sectoral wage differentials, labor mobility and the ratio of unemployment to job vacancies in detailed studies of seven countries with a variety of labor market and macroeconomic structures: the United States, Japan, West Germany, Sweden, the United Kingdom, Italy and Spain.

40243-3 Hardcover \$64.50

Now in a new edition...

The Stages of Economic Growth

A Non-Communist Manifesto

Third Edition

W. W. Rostow

Reviews of earlier editions:

"...the most stimulating contribution to political and economic discussion made by any academic economist since the war."
—*The Economist*

"Imaginative, stimulating statement of the economic goals of technologically underdeveloped nations, and how they can be most effectively achieved, without resort to Communism."

—*The New York Times*

"This interesting, well-written and important book projects a new light on various problems and will be much discussed. Its 167 pages of text provide a world history of the last century or two in terms of the stages of economic growth of the principal nations." —*Financial Times*

40070-8 Hardcover \$47.50

40928-4 Paper \$16.95

Available in bookstores or write:

**CAMBRIDGE
UNIVERSITY PRESS**

40 West 20th Street, New York, NY 10011-4211.

Call toll-free 800-872-7423.

MasterCard/VISA accepted. Prices subject to change.

GENERAL EQUILIBRIUM MODELS FOR MICRO-COMPUTERS

GEMODEL.USA

A large-scale general equilibrium model of the U.S. economy.

The package is completely menu-driven and accomplishes every job from data entry through consistency checks, calibration and policy changes to solution and report printing.

The model can be used to analyze the effects of changes in tax policy, product demands, technology, productivity etc. on incomes, outputs, tax revenue and foreign trade.

PRICES:

Academic Version	US\$ 2,200
Professional Version	US\$ 8,100

GEMODEL.USA is for policy analysis and research in government, university and industry.

A special purpose of the model is to simulate value-added taxes and the separate tax policies of two levels of government:
Federal
State/Local

GEMODEL.USA has a phenomenal number-crunching capability. It accepts the 85-industry, U.S. input-output tables prior to aggregation of industries, consumption categories, and households. The level of detail carried by the model is larger than that commonly found in the general equilibrium literature.

The package is easy to run and is perhaps the only software that produces simulation results on your own desk within seconds. It is available in two versions: Professional and Academic.

The Professional version allows greater disaggregation and the computation of dynamically sequenced equilibria.

Both models are supplied with sample data that can be edited or replaced with your own or another country's data.

SYSTEM REQUIREMENTS: IBM-PC or compatible, 384K RAM, 1 drive, DOS 2.1 or better.

Available on 5¼ and 3½ inch diskettes.

GEMODEL 2.0

Solves open or closed, two or three-industry small open economy models employing two or three factors. Industries can have constant or diminishing returns to scale. Admits one to nineteen household groups with or without income/leisure choice. Solves with a wide array of factor, commodity and income taxes.

"GEMODEL is a powerful microeconomic simulation model."

"I highly recommend it for both instructional and research purposes." *Social Science Micro Computer Review*, 1987.

"This package opens up exciting possibilities for students to explore and verify economic theory..." *Economic Journal*, 1990.

PRICE: US\$ 395.00

GESTATS

Calibrates GEMODEL 2.0 parameters to data and elasticity assumptions. Checks data consistency.

PRICE: US\$ 195.00

GEREPORT

Compares GEMODEL 2.0 simulation results with the base case or with other simulation results. Computes measures of welfare change and marginal costs of taxes.

PRICE: US\$ 95.00

GEDATA

24 data files for exercises in price theory, trade and finance using GEMODEL 2.0

PRICE: US\$ 25.00

For information and to order, write to:

DIA Agency Inc., 1879 Kingsdale Ave., Ottawa, Ontario, K1T 1H9 CANADA
DEMONSTRATOR diskettes available at \$25.



Announcing four new versions of MicroTSP...

MicroTSP 7.0 for PC Compatibles

The best selling econometric software is now even better. New features include a wide range of model evaluation & hypothesis tests, recursive estimation, page layout features in graphics, forecast confidence intervals, and much more.

MicroTSP 6.6 for the Macintosh

This all new version was designed to make full use of the Apple Macintosh. View and manipulate multiple windows of statistical output and graphics using the standard Macintosh interface. Data capacity is limited only by available memory.

Student Versions for PC and Macintosh

These limited data capacity versions include all of the econometric and graphic techniques of MicroTSP 6.5. Price is comparable to the average textbook.

Standard Features of All Versions of MicroTSP

Menu, command, or batch operation • linear and nonlinear estimation of equations & systems • Box-Jenkins • ARMAX • exponential smoothing • logit • probit • PDLs • VAR • forecasting • model simulation • handling of missing data • extensive high resolution graphics • data transfer • and much more.

Write or call for more information and a free demonstration disk. Academic and quantity discounts are available.

**MICRO
TSP™**

Quantitative Micro Software
4521 Campus Drive, #336, Irvine, CA 92715
(714) 856-3368 • FAX (714) 856-2044

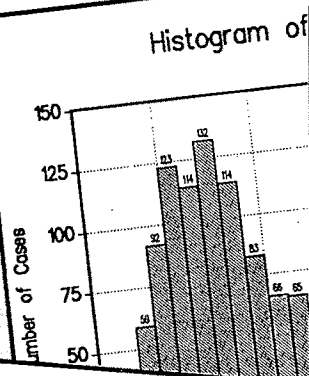
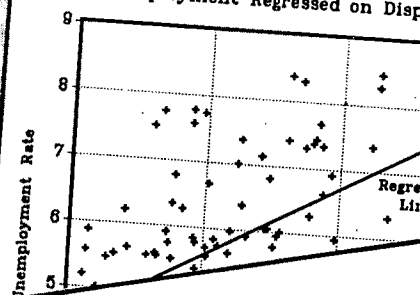


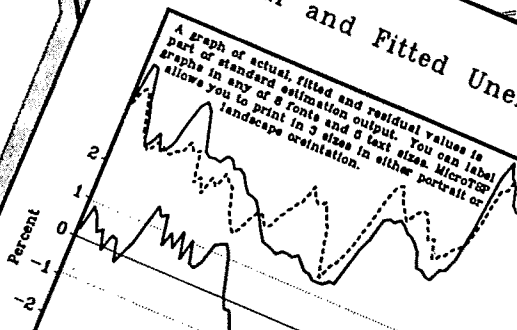
Figure 1
Unemployment Regressed on Dis



LS // Dependent Variable is UNEMP
Date: 1-10-1990 / Time: 15:36
SNPL range: 1958.1 - 1981.4
Number of observations: 96
Convergence achieved after 5 iterations

VARIABLE	COEFFICIENT	STD. ERROR	T-STAT.
C	2.6289650	1.7789567	1.4778129
TIME	0.0314951	0.0172281	1.8281233
SIG	1.5555103	1.0546551	1.4748995
SIG(-1)	2.1020108	1.1382556	1.8466949
SIG(-3)	-2.1029372	1.0998758	-1.9119769
DMR	-8.0349240	6.2269175	-1.2903534
DMR(-1)	-4.9759800	9.6008864	-0.5182834
DMR(-3)	-17.303940	10.737224	-1.6115841
DMR(-5)	-24.481701	9.6090656	-2.5477713
DMR(-7)	-16.981250	6.1537754	-2.7594848
DMR(-9)		0.0901606	17.294585
DMR(-11)		0.0911986	-6.8932053

Mean of dependent
D. of dependent
of squared res
-cic





Crane Russak

a Member of Taylor & Francis Group

Economic Policy Alternatives for the Latin American Crisis

Joan B. Anderson, Department of Economics,
University of San Diego, CA

This book explores macroeconomic policy alternatives available to Latin American policy-makers from both a theoretical and empirical perspective. It presents a quantitative framework in which to evaluate the effectiveness of various types of monetary and fiscal policies under the current conditions of inflation, declining growth, and debt.

1990 • 120 pages

0-8448-1677-9 Hardcover \$40.00

0-8448-1660-4 Softcover \$19.91

The Leadership Challenge of Economic Reforms in Africa

Edited by Olusegun Obasanjo, Africa Leadership Forum,
and Hans d'Orville, InterAction Council

This collection of statements by international financial public officials examines such diverse issues as Africa's economic, political and institutional problems as a prelude to effective economic reforms and international cooperation.

1991 • 112 pages

0-8448-1680-9 Hardcover \$34.00

FORTHCOMING!

Rethinking the Third World: Contributions Towards a New Conceptualization

Edited by Rosemary Galli, Consultant,
United Nations Development Program, US Agency for International
Development, and the EEC

1991 • 280 pages

0-8448-1711-2 Hardcover \$49.50

0-8448-1712-0 Softcover \$21.95

To Order, Call, TOLL-FREE, 1-800-821-8312
Or Write to: Crane Russak, c/o Taylor & Francis,
1900 Frost Road, Suite 101, Bristol, PA 19007-1598



The International Debt Crisis

Debt Equity Conversion - A Guide for Decision Makers

- *definition of policy objectives*
- *analysis and clarification of instrumentalities*
- *examination of national experiences through case studies*
- *assessment of creditor country constraints*

Debt Equity

CONVERSION

A Guide for
Decision-makers

In 1984-1989 alone, debt equity conversions amounted to \$22 billion. As substantial as this is, this new study concludes that these conversions have had limited impact on debt reduction and the creation of new investment. This thorough examination of a complex topic provides decision makers with the tools to define their options and take appropriate action in the area of debt equity conversion. This book reviews the basics of the debt equity concept, its history, and typical deals and examines the role of all the parties involved in conversions as well as the benefits to debtor countries and constraints on creditor countries. Notably, this guide offers a detailed checklist of issues that any country considering a debt equity conversion programme is obliged to confront. This study pays particular attention to the experiences of Mexico, Brazil, Chile, Argentina, the Philippines and Jamaica.

E.90.II.A.22 92-1-104352-2 \$27.50

Transnational Banks and the International Debt Crisis

Very little attention in this crisis is paid to the role of transnational banks (TNBs) even though a large part of the developing countries' debt is owed to them. This book analyses how TNBs contributed to the creation of the international debt problem by providing excessive loans in defiance of normal prudential criteria and how their subsequent activities have helped prolong the problem.

E.90.II.A.19 92-1-104349-2 \$22.50

International Debt Restructuring: Substantive Issues and Techniques

As a response to numerous requests from decision makers of debt restructuring, this study sets out the major substantive and legal issues needed to be addressed by sovereign borrowers facing the necessity of rescheduling their public external debt.

E.89.II.A.10 92-1-104317-4 \$10.00

Send orders to:

United Nations Publications, Sales Section, Room DC2-0853 Dept. 675
New York, N.Y. 10017 Tel. (800) 253-9646, (212) 963-8302, Fax. (212) 963-3489.
Visa, MasterCard and American Express accepted for orders over \$15.00.
Please add 5%, \$3.50 minimum for shipping & handling.

United Nations Publications

ySTAT / yCHART / yMED™

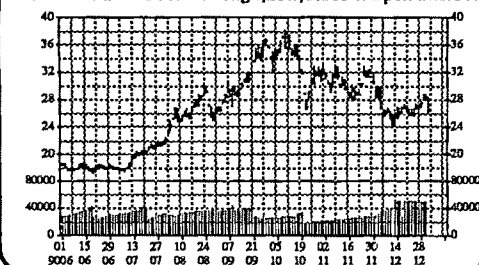
Integrates

SPREADSHEET + STATISTICS + FORECASTING + GRAPHICS

For all models of IBM PC's and Macintosh Computers

- ySTAT is the only full-fledged spreadsheet statistical program - a fast, powerful, easy-to-use stand-alone program.
- Menu-driven format as in Lotus 1-2-3 - offers many functions that are not available even in 1-2-3, such as vector formulas, lagged variables, dummy variables, moving averages, and interpolation of missing values.
- You can read, enter, edit, copy, move, and sort the data as well as group and partition the data for analysis right on the spreadsheet. A total of 40 function keys for one-keystroke operations.
- ySTAT reads 1-2-3 and Symphony worksheet files directly, and any tree-format or fixed-format textfile generated by other PC programs or downloaded from mainframe. The data can be numeric, alphabetic, including missing values.
- Mean, median, quartiles, skewness, mode, t-test, frequency distribution, correlation, crosstabs, analysis of variance.
- Multiple regression (OLS), standardized regression, weighted least squares, two-stage least squares, polynomial regression, and Cochran-Orcutt method. Including diagnostic residual analysis output (see sample at right).
- Pooling of cross-section and time-series data.
- Two-sample difference of means test, nonparametric tests (Kolmogorov-Smirnov test, Wald-Wolfowitz runs test, Mann-Whitney test, Wilcoxon matched-pairs signed-ranks test) and Spearman's and Kendall's rank-order correlations.
- Time-series models: autoregressive model, moving average model, and Box-Jenkins model (ARIMA).
- Nonseasonal forecasts: linear trend, quadratic trend, polynomial model, sinusoidal model; simple, double and triple exponential smoothings.
- Seasonal forecasts: exponential smoothings with seasonal indicators or with trigonometric functions; Winters' additive and multiplicative procedures.
- Logit, Probit, Tobit, and Weighted Probit regressions: these models are fast and easy to use.
- Will automatically use expanded memory (EMM) if available.
- High-resolution color charts: the actual and forecast values are graphically displayed as a line chart and also as an XY chart, including a regression line and its 95% confidence interval.

Crude Oil Futures Prices - High, Low, Close & Open Interest



- yCHART offers multiple charts in one window. There are line, XY, bar, stacked-bar, area, high-low-close-volume, pie, and survival charts. The Macintosh version has the addition of horizontal bar, histogram, box, moving average/oscillator, and spread charts.
- yMED is a medical version of ySTAT. It has, among others, Mantel-Haenszel odds ratio, risk difference and ratio, incidence rate difference and ratio; loglinear model; Kaplan-Meier survival model.
- APPLE MACINTOSH VERSION**
The Mac version builds on the features of the IBM PC version and extends the worksheet to three dimensional - there can be a maximum of 24 worksheets in a multiple-window environment. Each worksheet represents a separate data set; you can perform same or different statistical procedures on different worksheets displayed simultaneously on the screen. A series of worksheets can also be combined in several ways - appended, merged, combined (summed up). You can also copy, cut, move, and paste between worksheets to exchange data. The Mac yCHART offers many chart types for commodity futures trading.

IBM is a registered trademark of International Business Machines Corp.; Macintosh of Apple Computer, Inc.; 1-2-3 of Lotus Development Corp.

OLS -- DEPENDENT VARIABLE: Consump				
RIGHT-HAND VARIABLE	ESTIMATED COEFFICIENT	STANDARD ERROR	T-STATISTIC	PROB.
1 Profit	0.192934311	0.031211	T= 2.11527	0.049
2 P-1	0.089884898	0.030659	T= 0.93158	0.335
3 W/M	0.796218750	0.039944	T= 19.93342	0.000
4 Constant	16.23600212	1.362181	T= 12.46382	0.000
SUM OF SQUARED RESIDUALS = 17.879449 (DF=17)				
VARIANCE (MSE) = 1.051722				
STANDARD ERROR (ROOT MSE) = 1.025546				
R-SQUARED = 0.281908				
ADJUSTED R-SQUARED = 0.977657				
F-STATISTIC (3, 17) = 292.707555 (p=0.0004)				
SUM OF RESIDUALS = 0.020000				
DURBIN-WATSON STATISTIC = 1.367474				
Analysis of Variance				
	Source	SUM SQ	DF	MEAN SQ
	Due to Regression	923.550	3	307.850
	Residual	17.879	17	1.052
	Total	941.430	20	47.071

RESIDUAL ANALYSIS - Mean: -0.000 Adj. RMSE: 0.946 Mean Abs. % Err: 1.283

AUTOCORRELATIONS				
LAG	COEF	T-VAL		
1	0.181	0.43		
2	-0.059	-0.27		
3	-0.033	-0.15		
4	-0.037	-0.16		
5	-0.411	-1.81		
Ljung-Box statistic (chi-square 4 DF): 6.698 (p=0.1527)				

PARTIAL AUTOCORRELATION				
LAG	COEF	T-VAL		
1	0.181	0.43		
2	-0.059	-0.27		
3	-0.033	-0.15		
4	-0.037	-0.16		
5	-0.378	-1.73		

ACTUAL versus FITTED VALUES AND RESIDUALS				
SEQ	Actual	Fitted	Residual	Std. Err
1	41.98	42.22	-0.24	1.026
2	45.00	46.25	-1.25	1.026
3	49.20	50.77	-1.56	1.026
4	50.00	51.07	-1.07	1.026
5	52.60	52.59	0.00	1.026
6	55.10	54.22	0.88	1.026
7	56.20	54.22	1.98	1.026
8	57.30	56.25	1.05	1.026
9	57.80	58.29	-0.49	1.026
10	58.00	58.29	-0.29	1.026
11	58.40	58.29	0.11	1.026
12	58.60	58.29	0.31	1.026
13	58.70	58.29	0.41	1.026
14	58.70	58.29	0.41	1.026
15	58.70	58.29	0.41	1.026
16	58.70	58.29	0.41	1.026
17	58.70	58.29	0.41	1.026
18	58.70	58.29	0.41	1.026
19	58.70	58.29	0.41	1.026
20	58.70	58.29	0.41	1.026
21	58.70	58.29	0.41	1.026

LIST OF OUTLIERS - STANDARDIZED RESIDUALS GREATER THAN 1.50				
SEQ	Actual	Fitted	Residual	Std. Err
3	49.20	50.77	-1.56	0.967
16	57.70	56.08	1.616	0.958
21	69.70	71.87	-2.173	0.780

(Note: This sample output shows some of the options for regression diagnostics)

MING TELECOMPUTING INC.

23 Oak Meadow Road, P.O. Box 101, Lincoln Center
MA 01773, U.S.A. (617) 259-0391, (617) 259-1431; fax

- IBM version: () ySTAT for \$395. () yMED for \$395.
() yCHART for \$195. () ySTAT/yCHART for \$495.
() Trial disk for \$5. () Free information/sample output
- System: () IBM PC. () XT. () AT. () System 2 model____
() 386 PC. () 486 PC. () Other____
- DriveSize: () 5-1/2" 360KB. () 5-1/2" 1.2MB () 3-1/2"
- Options: () IBM Monochrome. () CGA. () EGA. () VGA.
() Hercules Mono. Card () Exp. Memory____MB
- Macintosh
Version: () yCHART for \$295. () ySTAT/yCHART for \$495.
() yMED/yCHART for \$495.
() Trial disk for \$5. () Free information/sample output
- System: () Plus. () SE. () SE/30. () II model____Memory____
- Payment: () Check or money order.
() U.S. university or governmental purchase order.
- () Visa. () MC. Card #____

Name____ Exp. Date ____/____/____

Address____

Telephone: ()____

Informative Additions to Economic Knowledge.

Economic Analysis of Industrial Policy

Motoshige Itoh, Kazuharu Kiyono,
Mashahiro Okuno-Fujiwara, and
Kotaro Suzumura

*A Volume in the ECONOMIC THEORY,
ECONOMETRICS, AND MATHEMATICAL
ECONOMICS Series (edited by Karl Shell)*

Written by four prominent Japanese economists, this book is a preliminary step for building a general theoretical framework for analyzing industrial policies. It achieves this goal by analyzing industrial policies in terms of economic theory, discussing the various frameworks suitable for evaluating these industrial policies in an objective manner, and by determining which policy measures are appropriate if an industrial policy is to be implemented.

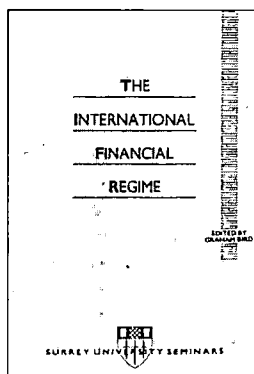
Key Features:

- Discusses a general framework for adopting industrial policy from the perspective of post-war Japanese economists
- Discusses problems facing developing countries, especially the newly industrializing countries, that adopt Japan's strategies
- Provides a theoretical analysis of Japanese industrial policies and their effect on economic welfare, taking into consideration their distinctive features

CONTENTS: (Section heads)

Industrial Promotion and Trade. Oligopolistic Control of an International Market. Welfare Implications of Strategic Competition in Oligopolistic Industries. Research and Development Investment. Further Problems in Formulation of Future Industrial Policy. References. Index.

June 1991, c. 280 pp., \$74.95 (tentative)
ISBN: 0-12-375735-5



Game Theory and Applications

edited by

Tatsuro Ichiishi,
Abraham Neyman,
and Yair Tauman

*A Volume in the ECONOMIC THEORY,
ECONOMETRICS, AND MATHEMATICAL
ECONOMICS Series (edited by Karl Shell)*

Key Features:

- Highlights new research directions in economic theory which surpass the neoclassical paradigm
- Includes game-theoretical analyses in economics, political science, and biology
- Written by leading game theorists, economists, political scientists, and biologists

December 1990, 436 pp., \$89.95

ISBN: 0-12-370182-1

*Order from your local bookseller
or directly from*



ACADEMIC PRESS

Harcourt Brace Jovanovich, Publishers
Book Marketing Department #01061
1250 Sixth Avenue, San Diego, CA 92101

CALL TOLL FREE

1-800-321-5068

FAX **1-800-235-0256**

Quote this reference number for free postage and
handling on your prepaid order **#01061**

Prices subject to change without notice.
© 1991 by Academic Press, Inc. All Rights Reserved. MJ/D/SS—01061.

The International Financial Regime

edited by

Graham Bird

*A Volume in the SURREY SEMINARS
IN ECONOMICS Series*

A Co-publication with Surrey University Press

International Financial Regime

addresses the central questions of international finance and economics in a lively and readable style. The contributors are all acknowledged experts in their fields and each contribution presents an accessible survey of the often complex issues which make up the international monetary system. This book is a useful reference for economists, policy makers, and advanced students of international economics.

1990, 360 pp., \$75.00

ISBN: 0-12-099745-2

Trade, Policy, and International Adjustments

edited by

Akira Takayama,
Michihiro Ohyama,
and Hiroshi Ohta

*A Volume in the ECONOMIC THEORY,
ECONOMETRICS, AND MATHEMATICAL
ECONOMICS Series (edited by Karl Shell)*

Key Features:

- Examines recent issues in the theory of international trade and its application to commercial policies, such as voluntary export restrictions, trade versus aid, and multinational corporations
- Analyzes macroeconomic issues such as exchange rates, monetary economies, and tariffs and quotas
- Extends trade theory to investigate the implications of variable returns to scale, intermediate inputs, and the question of choice between free and controlled trade

February 1991, 305 pp., \$45.00

ISBN: 0-12-682230-1

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

announces
a special issue in honor of Professor A. SEN

ALTERNATIVES TO WELFARISM

Volume 56 (3-4) 1990

Edited by M. De Vroey

Contents:

- M. DE VROEY, Introduction
- Ph. VAN PARIJS, Academic Presentation of Amartya Sen
- Ch. BLACKORBY, D. DONALDSON and J. WEYMARK, A Welfarist Proof of Arrow's Theorem
- J. ROEMER, Welfarism and Axiomatic Bargaining Theory
- H. MOULIN, Interpreting Common Ownership
- Ph. VAN PARIJS, Equal Endowments as Undominated Diversity
- G.A. COHEN, Equality of What? On Welfare, Goods and Capabilities
- P. PATTANAIK and Y. XU, On Ranking Opportunity Sets in Terms of Freedom of Choice
- H. STEINER, Putting Rights in their Place: An Appraisal of A. Sen's Work on Rights
- Fr. BOURGUIGNON and G. FIELDS, Poverty Measures and Anti-Poverty Policy
- E. SCHOKKAERT and L. Van OOTEGEM, Sen's Concept of the Living Standard Applied to the Belgian Unemployed
- A. SEN, Welfare, Freedom and Social Choice: A Reply

Name:

ADDRESS:

☐ Please send me a copy of the special issue "**Alternatives to Welfarism**",
(40 US\$, 1000 FB)

☐ Please enter my subscription to **Recherches Economiques de Louvain**
(70 US\$, 2000 FB)

☐ Payment enclosed:

☐ Cheque (made available to De Boeck Wesmael S.A.)

☐ Credit Card : ☐ Visa ☐ Master Card ☐ American express

Expiry date Card number Signature

☐ Please send me an invoice.



Send your order to:

DE BOECK WESMAEL S.A.

Avenue Louise, 203, boîte 1 - B-1050 Bruxelles - Belgium

Phone: 32 / 2 / 640 72 72 - Fax: 32 / 2 / 641 92 84

JOB OPENINGS FOR ECONOMISTS

A bi-monthly listing of academic and nonacademic jobs open at press time. Published seven times a year—February, April, June, August, October, November (a supplement to October), and December.

Annual Subscription Rates

	USA, Canada & Mexico (First Class)	Foreign (Air Mail)
AEA Junior Members	\$ 7.50	\$15.00
AEA Regular Members	\$15.00	\$22.50
Non-members & Institutions	\$25.00	\$32.50

Please begin my issues with:

☐ February ☐ April ☐ June ☐ August ☐ October ☐ December

FIRST NAME AND INITIAL															LAST NAME															SUFFIX				
ADDRESS LINE 1, OR ATTENTION																																		
ADDRESS LINE 2																																		
CITY															STATE OR COUNTY															ZIP/POSTAL CODE				

PLEASE TYPE OR PRINT INFORMATION ABOVE; DO NOT EXCEED SPACES ALLOWED.

PLEASE SEND CHECK OR MONEY ORDER PAYABLE IN U.S. DOLLARS. CANADIAN AND FOREIGN PAYMENTS MUST BE IN THE FORM OF A CHECK DRAWN ON A NEW YORK BANK PAYABLE IN U.S. DOLLARS.

Please send with payment to:

AMERICAN ECONOMIC ASSOCIATION
P.O. Box 307026
NASHVILLE, TENNESSEE 37230-7026
U.S.A.

INDEX OF ECONOMIC ARTICLES

prepared by

The Journal of Economic Literature
of the
American Economic Association

- ✓ Each volume in the **Index** lists articles in major economic journals and in collective volumes published during a specific year.
- ✓ No other single reference source covers as many articles classified in economic categories as the **Index**.

Index volumes XI-XXII covering 1969-1980 are available at \$60.00 each.

The following two part **Index** volumes are now ready for delivery at \$90.00 per set:

Volume	Year
XXIII	1981
XXIV	1982
XXV	1983
XXVI	1984
XXVII	1985
XXVIII	1986
XXIX	1987

*an
indispensable
tool for...*

ECONOMISTS
REFERENCE LIBRARIANS
RESEARCHERS
TEACHERS
STUDENTS
AUTHORS

*Note: 25% discount for AEA members
(individuals only).*

Payment required in advance. Prices include shipping charges; allow 4-6 weeks for delivery. Please send your check or money order (net of applicable discount) payable in United States dollars drawn on a United States bank to:

American Economic Assn. - **Index**
P.O. Box 307026
Nashville, TN 37230-7026

Address inquiries or other correspondence to:
Journal of Economic Literature, P.O. Box 7320
Oakland Station, Pittsburgh, PA 15213.

The Foremost Economic Literature Database!

Published by the
American Economic Association

Online

Economic Literature Index

File 139 on DIALOG®

In-depth coverage of journal literature: citations from the Journal of Economic Literature since 1969, with abstracts since 1984. Also citations of articles in collective volumes since 1979.

On Disc

EconLit

CD-ROM by SilverPlatter®

Over two decades of citations and abstracts from the Journal of Economic Literature on a single disc! Includes abstracts of journal articles and books, citations of articles in collective volumes, and dissertation titles.

In Print

Index of Economic Articles in Journals and Collective Volumes

Over 100 years of economic literature indexed by publication year: 1886 through 1987 in 29 volumes on your reference shelf! Make sure your set is up-to-date.

For information, write or call:

Journal of Economic Literature, P.O. Box 7320, Pittsburgh, PA
15213-0320 (412) 268-3869

SEOUL JOURNAL OF ECONOMICS: Call for Papers



Editors

Jae-Yoon Park
Shin-Haing Kim
Seoul National University

Associate Editors

Kwan Koo Yun
SUNY at Albany
Jaymin Lee
Yonsei University

Board of Editors

Hyung-Yoon Byun
Byung-Kwon Cha
Seoul National University
Gary S. Becker
University of Chicago
Jean C. Bénard
University of Paris I
James A. Mirrlees
Oxford University
Takashi Negishi
University of Tokyo
Jeffrey D. Sachs
Harvard University

in association with Institute
of Economic Research,
Seoul National University

Director

Byong-Jick Ahn

Seoul Journal of Economics is designed to provide an outlet for research on all aspects of economic development, especially focusing on the economic development of East Asia. The journal also seeks to publish recent theoretical developments in all specialized fields of economics as well as rigorous empirical works.

Some of the papers featured in the first three volumes are

Jeffrey D. Sachs

Prospects for Global Trade Imbalances: A Simulation Approach

Ronald I. McKinnon and Kenichi Ohno

Getting the Exchange Rate Right: Insular vs. Open Economies

Andrew Caplin and Kala Krishna

Tariffs and the Most-Favored-Nation Clause: A Game Theoretic Approach

Submission of papers (in triplicate) should be addressed to

Editors, *Seoul Journal of Economics*,
Institute of Economic Research,
Seoul National University,
Seoul 151-742, Korea

or Professor Kwan Koo Yun,
Dept. of Economics,
SUNY at Albany,
Albany NY 12222 (inside USA)

Annual subscription rate (4 issues):

☐ Individuals US \$20.00 ☐ Institutions US \$40.00

ORDER FORM

Payment is enclosed-Check () Money Order ()

Name: _____

Address: Street no. _____

City _____

State _____

To order please mail this coupon.

Make check or money order payable to

Seoul Journal of Economics

INSTITUTE OF ECONOMIC RESEARCH
SEOUL NATIONAL UNIVERSITY

Seoul 151-742, Korea

Tel : 82-2-877-1629 Fax : 82-2-888-4454